Rat genome variation and complex traits

Roel Hermsen

ISBN / EAN: 978-94-6203-696-3

Art work: Tom & Merel Schelland and Thijs Versloot

Printed by: CPI Koninklijke Wöhrmann

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

Rat genome variation and complex traits

Genoomvariatie in de rat en complexe eigenschappen

(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 21 januari 2015 des middags te 12.45 uur

door

Robertus Martinus Theodorus Hermsen

geboren op 23 januari 1984
te Huissen

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

Contents

1

Chapter 1

General introduction

1

Understanding human complex traits

The role of the laboratory rat in biomedical research has been pivotal over the last 150 years as a model in physiology, pharmacology, toxicology, nutrition, behavior, immunology and neoplasia [1]. In these fields the rat has been the preferred model organism, due to its size, ease of manipulation, learning capacities and breeding characteristics. Whereas for instance mouse, had just gained popularity when the knockout technologies flourished. In addition, the rat has been of great value for understanding the function of the vertebrate and human genomes. One way in which the intrinsic value of the rat as an animal model can be fully utilized is by coupling genetic elements to complex traits as a model for human complex diseases. The genetic basis of complex diseases is difficult to study in human due to the polygenic nature of these traits and the heterozygosity of the outbred human genome. The inbred rat strains that are used to model complex traits possess two fully identical copies of their genome. So using rat as a model for human complex disease allows the delicate dissection of the underlying genetic elements behind these phenotypes, whereas mouse has historically been used more for studying single gene phenotypes. To this extent, a broad diversity of rat strains have been bred over the past decades. Among these, models for complex diseases such as hypertension, type I diabetes, cancer and neurological disorders [2]. The development of NGS-based techniques has had big implications for biomedical research using the laboratory rat. With the emergence of next-generation sequencing (NGS) the rat genetic code and its function is slowly being unraveled. Here we review the impact, implications and challenges of NGS within the rat functional genomics field.

Studying genotype-phenotype in 'pre-reference-genome-times'

In the process of linking a phenotype to a specific locus in the genome, genetic markers are essential. The first steps in exploring the genome of the rat was by single gene and marker mapping in the early nineties. These markers could be used for the generation of phenotype specific quantitative trait loci (QTL) and disease gene mapping. In 1991, 214 genes had been assigned to chromosomes, most by analysis of a somatic cell hybrid panel [3]. The genetic linkage map, however, only consisted of 11 linkage groups spanning a relatively small fraction of the genome [3]. In that same year, 112 DNA marker polymorphisms were used to map

high blood pressure QTLs in the stroke-prone spontaneously hypertensive rat (SHRSP) [4]. Halfway the nineties several studies published newly identified markers per chromosome [5-12]. Nonetheless the overall genetic map still provided only partial coverage of the genome. This led to the construction of a more complete genetic linkage map of the laboratory rat in 1995 [13]. It consisted of 432 genotyped simple sequence length polymorphisms (SSLPs) and appeared to be linked to 99.5% of the rat genome. Because these markers show an average polymorphism rate of ~50% among common inbred strains, the map made it possible to scan the genome in most rat crosses. Over the following years marker numbers steadily increased from 767 in 1997 [14], about 5,000 in 1999 [15,16], to more than 24,000 in 2004 [17]. The increased availability of genetic markers has led to the identification of QTLs in several complex phenotypes such as blood pressure [18], diabetes [19-21], cardiovascular disease [22,23], stroke [24], ethanol preference [25], behavioral conditioning and anxiety [26,27], fat accumulation [28], arthritis [29] and chemical carcinogenesis [30]. Together with the progress in the mouse and human fields, the identified DNA markers made it possible to translate phenotypic QTLs between species identifying for instance regions in the human genome that are very likely to contain hypertension genes [31]. Taken together, the steady increase of genetic markers and insights, provided a solid basis for the construction of the first rat reference genome in 2003 [32].

Studying genotype-phenotype in 'pre-NGS-times'

Over the course of the last decades, (strain specific) genetic information became gradually available for rat researchers. A boost for the rat community was the publication of the first reference assembly of the Brown Norway rat in 2004 [32], just before the full emergence of NGS. In addition the STAR Consortium published three million newly detected SNVs based on six different strains [33]. Next to that they presented genotype information of more than 20,000 SNVs in 167 distinct inbred rat strains, two rat recombinant inbred panels and an F2 intercross. The availability of a reference genome in combination with strain specific genotype information provided the opportunity to carefully map previously identified genetic markers and use new genetic markers for QTL mapping. Also new types of data were used to map QTLs. Not solely the 'classical' phenotypic QTLs (phQTLs) were mapped, but also expression QTLs [34] or metabolomic QTLs (mQTLs) [35] were mapped to the rat genome,

integrating genetic data with metabolic [35] or array expression [34] profiles. Altogether the accumulated and improved genetic resources combined with new types of data resulted in the identification of additional QTLs for several complex traits including diabetes [35], hypertension [34], metabolic syndrome [36], depression [37], osteoporosis [38] and alcohol consumption [39]. QTL mapping is often followed by confirmation of the loci by the development of (consomic and) congenic lines, which contain a specific introgressed parts of DNA of one strain into another one [40] and which have successfully been used in the last decade to investigate human complex phenotypes such as cardiovascular disease [41] and type 2 diabetes [42].

The impact of next-generation sequencing

Genome sequencing

Nowadays the collection of different rat strains that cover a broad range of complex phenotypes and include crosses to fine map specific traits comprises more than 2800 strains [43]. The introduction of next-generation sequencing brought an enormous opportunity to more sensitively detect and molecularly dissect QTLs in all of these strains. One route is by the availability of whole-genome sequence (WGS) data and thus variant positions compared to the reference genome. The first step in this route was the publication of a genuine reference genome of the Brown Norway rat strain (BN/NHsdMcwi) in which NGS data was incorporated [32]. Since this assembly, multiple updates have been released, with the version 5.0 being the latest in 2012 [44]. This latter assembly includes data of extra large insert mate-pair NGS libraries [45] . This data contributed directly to improved contig scaffolding of the assembly. Furthermore, also the transcript annotation became more precise, reducing the number of transcripts with about a quarter from 39,549 in RGSC3.4 to 29,188 in RGSC5.0. In line with human sequencing projects were focus shifted from reference genome sequencing to population sequencing [46], more strains were whole-genome sequenced over the past few years [47-52]. The first variation catalog of a non-reference inbred strain was published in 2010, by sequencing of the genome of the spontaneously hypertensive rat (SHR) [47]. In addition, the SHR data were later combined with another non-reference strain, BN-*Lx*, resulting in a comprehensive catalog of variation between these two founders of the HXB/BXH recombinant inbred

(RI) panel [48]. This panel consists of 30 inbred lines that have been inbred for currently >80 generations, which have undergone extensive phenotyping at the physiological, behavioral and molecular levels [48]. Gathering whole genome sequence information from more strains allowed researchers to investigate strain-specific regions under artificial selection containing putative candidate genes underlying cardiovascular disease [49]. In addition whole genome sequence information of strains modeling specific human diseases like metabolic syndrome and rheumatoid arthritis was used to identify candidate variants for these traits [51]. Furthermore, three strains used to model metabolic syndrome were genomically dissected to identify the haplotype structure of these strains [52]. Also in mouse, WGS data became readily available over the past decade [53]. In addition, by intensive analysis of NGS data, detailed inventories have been presented on large structural genomic variants in part caused by transposable elements [54-56]. These studies showed also that accurate detection of these kind of events is still a big challenge. In addition a study in mice identified 843 QTLs of 100 traits in the mouse heterogeneous stock cross as a starting point for functional analysis [53]. In summary, the generation of whole genome sequence data by the introduction of NGS resulted in independent variant catalogs of about 40 widely used inbred rat strains.

Other NGS techniques

The introduction of NGS did not only have a high impact on genome sequencing. Also a broad range of other applications have been and are being developed over time. These applications all assay different cell compartments or cell-states, which allows the collection of very specific information on cell dynamics and even single cell metabolism. One of the largest applications besides genome sequencing, is the sequencing of RNA, which was first published in the same year as the first rat non-reference genome [47]. RNA sequencing allows researcher to assay a given transcriptome in an unbiased annotation-independent way, as the design of an expression array relies on the quality of the annotated genome. In this way hypothesis-free comparison of expression levels in two conditions during for instance treatment with toxicological agents can be measured [57], making it a broadly applicable technique in functional genomic research. Comparison of both platforms showed that RNA sequencing was more sensitive in detecting genes with low expression levels, while similar gene expression patterns were observed for both platforms [57]. Over

the past years, larger sample numbers have been studied [58] in part due to an optimized library preparation. With the availability of complete transcriptomes also questions on the regulation of transcription can now be challenged in a genome-wide fashion. Proteins interacting with DNA can mark several states of chromatin or initiation sites for transcription. These proteins and their modifications can be used to pool down and sequence the bound DNA segments (ChIP sequencing). With this technique the different functionalities of the noncoding genome can be explored, looking for instance at the influence of genomic variants on transcription factor binding [59-62]. The increasing interest in the noncoding part of the genome was underscored by the ENCODE consortium, which identified a broad inventory on DNA protein interactions in human tissues and cell lines [63]. Identification of functional noncoding genomic elements in rat by NGS started in 2011 by comparing genome-wide distribution of RNA polymerase III in six mammals using ChIP-seq [64]. In the following years more functional elements in the rat genome were identified by use of ChIP sequencing [65-67]. In addition to this 'direct' layer of epigenetic regulation, also 'indirect' epigenetic regulation was recently described by characterizing the 3D chromatin organization in the nucleus in mouse and human. This study identified large topological domains that show distinct features and are highly conserved among species [68]. For rat, this type of data has not been generated yet. But, looking at the progress and promise of this type of techniques, rat will probably quickly follow mouse and human. Altogether NGS techniques have been applied on a large scale from 2010 onwards, gaining fast insight in the loci and variants underlying complex traits. Next, exploring the function of (non)coding DNA sequence asks for validation methods for specific DNA variants.

Approaches in candidate variant validation

Besides the technological developments in the NGS field, also the techniques that are used to generate specific gene-knockout strains have improved significantly in the past decade. In part driven by the need of validation of found candidate variants from QTL mapping studies, precision tools to manipulate the rat genome have been developed to reproduce the phenotype of interest on a different genetic background. In general two approaches can be taken in generating a knockout allele for a specific gene. One way is by forward genetic approaches such as targeted selected ENU- or transposon-based mutagenesis [69,70]. In the ENU approach, random mutagenic processes damage the DNA of

spermatogonial stem cells in a treated male, which is then crossed with untreated females to create heterozygous carriers for each mutation. Loci of interest can then be screened for the presence of a mutation. Using this approach, rat researchers build ENU and transposon-based archives of mutants to facilitate further research [71]. Another way of generating specific knockout-strains is by forward genetic genome editing. The classical approach for this is homologous recombination (HR) in pluripotent ES cells [72]. The stable culturing of ES [73,74] and iPS [75,76] cells lines allowed the application of HR in rat for the first time in 2010 by targeting of the *Tp53* gene [77]. The authors of this work also provided the rat community with a detailed protocol [78] to apply HR in rat ES cells. In addition, independent of pluripotent stem cells, techniques with increasing specificity and decreasing off-target effects were developed. Three techniques have been applied in rat for the first time in the past five years: Zinc-finger nucleases (ZFNs) in 2009 [79], TAL (Transcription activator-like) effector nucleases (TALENs) in 2011 [80] and CRISPR-Cas in 2013 [81]. All are based on the same principle: design of a specific construct that results in a double-strand DNA break (DSB) which is error-prone repaired by nonhomologous end joining (NHEJ). In the coding region of a gene of interest this can lead to out-of-frame mutations due to insertion or deletion of a few bases by NHEJ. In addition, the CRIPR-Cas system allows presentation of recombinant homologues DNA, which facilitates insertions or deletions. These tools can efficiently manipulate the rat genome, even in an allele- [82] or tissue-specific manner [83], and thus have the potential to validate previous findings in QTL mappings studies and generate rat models with human disease genes or variants.

Challenges and outlook

Next-generation sequencing has proven its value in biomedical research including the rat as an animal model in the past decade. After the introduction of these techniques, work followed the path taken in human functional genomics field: from sequencing and assembly of a reference genome, the identification of variation between individual strains/individuals, to the characterization of differences between cells within an individual strain [46]. In rat the first two steps are in part taken: a reference assembly has been available since 2003 and whole-genome sequence of about 40 strains is now readily accessible [47-52]. Since the large number of rat strains that are used in functional genomics research, WGS data generation will probably continue over the coming years. A

challenge in the continuous production of WGS data for more strains is the concordance in variant calling algorithms and use of one version of the reference genome. Simultaneous analysis with a single variant caller and reference genome will be essential to accurately interrogate one specific locus in WGS data from several studies. In addition to the marker panels, WGS data can now be used to identify QTLs for a diversity of complex traits. Although QTL identification has already been performed using existing marker mapping panels, a given QTL still contains multiple genes and is in general considered to be rather large. An additional round of congenic breeding is then necessary to narrow down one specific QTL in size, which requires more animals and is both expensive and time consuming. Another challenge is the interpretation of noncoding DNA sequence. The identification of regions that initiate or regulate transcription is now becoming more common, but linking them directly to a polygenic trait is and will be difficult. In line with human functional genomics, also the characterization of differences between cell-types within an individual strain has been kicked off. For instance by the running FP7 funded EURATRANS consortium (2010-2015), which aims to assay RNA and epigenetic modifications in heart and liver tissue in a subset of strains for which genomic data was already available [84]. The overall characterization of cell-type specific differences will continue throughout the next decade. With this information we will get more insight into the heritable component of complex diseases. Still the etiology of many of these diseases, such as hypertension and cancer, is not fully understood and remains to be elucidated. Besides the biological challenges for the rat functional genomics field, the ease of generating NGS based data sets forces the community to overcome the ´data dogma´ in biology. Systematic integration of different data modalities is pulling biologists out of their comfort zone and requires the development of new methods that overcome the linear idea on DNA » RNA » protein by multidimensional modeling. One of the first steps is integrating just two data modalities, which was demonstrated recently by integrating RNA and protein data [85]. In addition meta-analysis on RNA sequencing data can identify co-expression networks modules and hubs [86], which are left uncovered by typical case-control RNA-seq experiments. These two examples show that bringing in more data types requires data expertise and tenacity.

In conclusion, the introduction of next-generation sequencing has brought a revolution in (not only) the rat functional genomics field by allowing researchers to carefully detect and dissect complex disease QTLs.

Outline of this thesis

The work in this thesis describes the identification and exploration of functional genomic variants in relation to complex traits in a variety of rat strains.

Chapter 2 describes the concordant identification of genomic variants in 40 rat strains and the characterization of substrain variants. I show that substrains harbor genomic variants with a relatively high impact, which may influence cellular processes.

In **Chapter 3** I discuss the use of whole genome sequencing data of the heterogeneous stock progenitor strains to impute the genomes of the 1407 NIH-HS rats and map QTLs. This work identified 355 high resolution QTLs for 122 phenotypes and serves as basis for further molecular dissection.

To assess the function of noncoding variants, I describe in **Chapter 4** an integrative genomics approach to study the effect of genetic variation in different rat backgrounds on the chromatin state and nuclear organization in liver tissue from ten inbred rat strains.

In **Chapter 5** I describe the molecular characterization of tumors from rat *Tp53* knockout model, generated in a targeted-selected ENU screen. Here, I try to find clues for the observed stable genomes in the homozygous knockout animals.

Finally, in the **General discussion**, I describe the possible challenges and outlook for the rat functional genomics field. Furthermore I discuss the implications of the use of next-generation sequencing techniques from a public perspective and the genetic awareness of NGS in society.

References

1. Jacob, H.J. (1999) Functional genomics and rat models. Genome research, 9, 1013-1016.
2. Smits, B.M. and Cuppen, E. (2006) Rats go genomic. Genome biology, 7, 306.
3. Levan, G., Szpirer, J., Szpirer, C., Klinga, K., Hanson, C. and Islam, M.Q. (1991) The gene map of the Norway rat (Rattus norvegicus) and comparative mapping with mouse and man. Genomics, 10, 699-718.
4. Jacob, H.J., Lindpaintner, K., Lincoln, S.E., Kusumi, K., Bunker, R.K., Mao, Y.P., Ganten, D., Dzau, V.J. and Lander, E.S. (1991) Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. Cell, 67, 213-224.
5. Cash, J.M., Remmers, E.F., Goldmuntz, E.A., Crofford, L.J., Zha, H., Hansen, C.T. and Wilder, R.L. (1993) Genetic mapping of the athymic nude (RNU) locus in the rat to a region on chromosome 10. Mammalian genome : official journal of the International Mammalian Genome Society, 4, 37-42.
6. Goldmuntz, E.A., Remmers, E.F., Zha, H., Cash, J.M., Mathern, P., Crofford, L.J. and Wilder, R.L. (1993) Genetic map of 12 polymorphic loci on rat chromosome 1. Genomics, 16, 761-764.
7. Remmers, E.F., Du, Y., Zha, H., Goldmuntz, E.A. and Wilder, R.L. (1995) Ten polymorphic DNA loci, including five in the rat MHC (RT1) region, form a single linkage group on rat chromosome 20. Immunogenetics, 41, 316-319.
8. Remmers, E.F., Goldmuntz, E.A., Cash, J.M., Zha, H., Crofford, L.J., Misiewicz-Poltorak, B., Mathern, P. and Wilder, R.L. (1993) Map of seven polymorphic markers on rat chromosome 14: linkage conservation with human chromosome 4. Mammalian genome : official journal of the International Mammalian Genome Society, 4, 90-94.
9. Remmers, E.F., Goldmuntz, E.A., Zha, H., Crofford, L.J., Cash, J.M., Mathern, P., Du, Y. and Wilder, R.L. (1993) Linkage map of seven polymorphic markers on rat chromosome 18. Mammalian genome : official journal of the International Mammalian Genome Society, 4, 265-270.
10. Remmers, E.F., Goldmuntz, E.A., Zha, H., Mathern, P., Du, Y., Crofford, L.J. and Wilder, R.L. (1993) Linkage map of nine loci defined by polymorphic DNA markers assigned to rat chromosome 13. Genomics, 18, 277-282.
11. Yamada, J., Kuramoto, T. and Serikawa, T. (1994) A rat genetic linkage map and comparative maps for mouse or human homologous rat genes. Mammalian genome : official journal of the International Mammalian Genome Society, 5, 63-83.
12. Zha, H., Wilder, R.L., Goldmuntz, E.A., Cash, J.M., Crofford, L.J., Mathern, P. and Remmers, E.F. (1993) Linkage map of 10 polymorphic markers on rat chromosome 2. Cytogenetics and cell genetics, 63, 117-123.
13. Jacob, H.J., Brown, D.M., Bunker, R.K., Daly, M.J., Dzau, V.J., Goodman, A., Koike, G., Kren, V., Kurtz, T., Lernmark, A. et al. (1995) A genetic linkage map of the laboratory rat, Rattus norvegicus. Nature genetics, 9, 63-69.
14. Bihoreau, M.T., Gauguier, D., Kato, N., Hyne, G., Lindpaintner, K., Rapp, J.P., James, M.R. and Lathrop, G.M. (1997) A linkage map of the rat genome derived from three F2 crosses. Genome research, 7, 434-440.
15. Steen, R.G., Kwitek-Black, A.E., Glenn, C., Gullings-Handley, J., Van Etten, W., Atkinson, O.S., Appel, D., Twigger, S., Muir, M., Mull, T. et al. (1999) A high-density integrated genetic linkage and radiation hybrid map of the laboratory rat. Genome research, 9, AP1-8, insert.
16. Watanabe, T.K., Bihoreau, M.T., McCarthy, L.C., Kiguwa, S.L., Hishigaki, H., Tsuji, A., Browne, J., Yamasaki, Y., Mizoguchi-Miyakita, A., Oga, K. et al. (1999) A radiation hybrid map of the rat genome containing 5,255 markers. Nature genetics, 22, 27-36.
17. Kwitek, A.E., Gullings-Handley, J., Yu, J., Carlos, D.C., Orlebeke, K., Nie, J., Eckert, J., Lemke, A., Andrae, J.W., Bromberg, S. et al. (2004) High-density rat radiation hybrid maps containing over 24,000 SSLPs, genes, and ESTs provide a direct link to the rat genome sequence. Genome research, 14, 750-757.
18. Garrett, M.R., Saad, Y., Dene, H. and Rapp, J.P. (2000) Blood pressure QTL that differentiate Dahl salt-sensitive and spontaneously hypertensive rats. Physiological genomics, 3, 33-38.
19. Galli, J., Li, L.S., Glaser, A., Ostenson, C.G., Jiao, H., Fakhrai-Rad, H., Jacob, H.J., Lander, E.S. and Luthman, H. (1996) Genetic analysis of non-insulin dependent diabetes mellitus in the GK rat. Nature genetics, 12, 31-37.
20. Jacob, H.J., Pettersson, A., Wilson, D., Mao, Y., Lernmark, A. and Lander, E.S. (1992) Genetic dissection of autoimmune type I diabetes in the BB rat. Nature genetics, 2, 56-60.
21. Pravenec, M., Gauguier, D., Schott, J.J., Buard, J., Kren, V., Bila, V., Szpirer, C., Szpirer, J., Wang, J.M., Huang, H. et al. (1996) A genetic linkage map of the rat derived from recombinant inbred strains. Mammalian genome : official journal of the International Mammalian Genome Society, 7, 117-127.
22. Stoll, M., Cowley, A.W., Jr., Tonellato, P.J., Greene, A.S., Kaldunski, M.L., Roman, R.J., Dumas, P., Schork, N.J., Wang, Z. and Jacob, H.J. (2001) A genomic-systems biology map for cardiovascular function. Science, 294, 1723-1726.
23. Moreno, C., Dumas, P., Kaldunski, M.L., Tonellato, P.J., Greene, A.S., Roman, R.J., Cheng, Q., Wang, Z., Jacob, H.J. and Cowley, A.W., Jr. (2003) Genomic map of cardiovascular phenotypes of hypertension in female Dahl S rats. Physiological genomics, 15, 243-257.

24. Rubattu, S., Volpe, M., Kreutz, R., Ganten, U., Ganten, D. and Lindpaintner, K. (1996) Chromosomal mapping of quantitative trait loci contributing to stroke in a rat model of complex human disease. Nature genetics, 13, 429-434.

25. Murphy, J.M., Stewart, R.B., Bell, R.L., Badia-Elder, N.E., Carr, L.G., McBride, W.J., Lumeng, L. and Li, T.K. (2002) Phenotypic and genotypic characterization of the Indiana University rat lines selectively bred for high and low alcohol preference. Behavior genetics, 32, 363-388.

26. Fernandez-Teruel, A., Escorihuela, R.M., Gray, J.A., Aguilar, R., Gil, L., Gimenez-Llort, L., Tobena, A., Bhomra, A., Nicod, A., Mott, R. et al. (2002) A quantitative trait locus influencing anxiety in the laboratory rat. Genome research, 12, 618-626.

27. Flint, J. (2003) Analysis of quantitative trait loci that influence animal behavior. Journal of neurobiology, 54, 46-77.

28. Tanomura, H., Miyake, T., Taniguchi, Y., Manabe, N., Kose, H., Matsumoto, K., Yamada, T. and Sasaki, Y. (2002) Detection of a quantitative trait locus for intramuscular fat accumulation using the OLETF rat. The Journal of veterinary medical science / the Japanese Society of Veterinary Science, 64, 45-50.

29. Olofsson, P., Holmberg, J., Pettersson, U. and Holmdahl, R. (2003) Identification and isolation of dominant susceptibility loci for pristane-induced arthritis. Journal of immunology, 171, 407-416.

30. De Miglio, M.R., Pascale, R.M., Simile, M.M., Muroni, M.R., Calvisi, D.F., Virdis, P., Bosinco, G.M., Frau, M., Seddaiu, M.A., Ladu, S. et al. (2002) Chromosome mapping of multiple loci affecting the genetic predisposition to rat liver carcinogenesis. Cancer research, 62, 4459-4463.

31. Stoll, M., Kwitek-Black, A.E., Cowley, A.W., Jr., Harris, E.L., Harrap, S.B., Krieger, J.E., Printz, M.P., Provoost, A.P., Sassard, J. and Jacob, H.J. (2000) New target regions for human hypertension via comparative genomics. Genome research, 10, 473-482.

32. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E. et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature, 428, 493-521.

33. Consortium, S., Saar, K., Beck, A., Bihoreau, M.T., Birney, E., Brocklebank, D., Chen, Y., Cuppen, E., Demonchy, S., Dopazo, J. et al. (2008) SNP and haplotype mapping for genetic analysis in the rat. Nature genetics, 40, 560-566.

34. Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V. et al. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nature genetics, 37, 243-253.

35. Dumas, M.E., Wilder, S.P., Bihoreau, M.T., Barton, R.H., Fearnside, J.F., Argoud, K., D'Amato, L., Wallis, R.H., Blancher, C., Keun, H.C. et al. (2007) Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. Nature genetics, 39, 666-672.

36. Bilusic, M., Bataillard, A., Tschannen, M.R., Gao, L., Barreto, N.E., Vincent, M., Wang, T., Jacob, H.J., Sassard, J. and Kwitek, A.E. (2004) Mapping the genetic determinants of hypertension, metabolic diseases, and related phenotypes in the lyon hypertensive rat. Hypertension, 44, 695-701.

37. Solberg, L.C., Baum, A.E., Ahmadiyeh, N., Shimomura, K., Li, R., Turek, F.W., Churchill, G.A., Takahashi, J.S. and Redei, E.E. (2004) Sex- and lineage-specific inheritance of depression-like behavior in the rat. Mammalian genome : official journal of the International Mammalian Genome Society, 15, 648-662.

38. Alam, I., Sun, Q., Liu, L., Koller, D.L., Fishburn, T., Carr, L.G., Econs, M.J., Foroud, T. and Turner, C.H. (2005) Whole genome scan for linkage to bone strength and structure in inbred Fischer 344 and Lewis rats. Journal of bone and mineral research : the official journal of the American Society for Bone and Mineral Research, 20, 1589-1596.

39. Tabakoff, B., Saba, L., Printz, M., Flodman, P., Hodgkinson, C., Goldman, D., Koob, G., Richardson, H.N., Kechris, K., Bell, R.L. et al. (2009) Genetical genomic determinants of alcohol consumption in rats and humans. BMC biology, 7, 70.

40. Flint, J., Valdar, W., Shifman, S. and Mott, R. (2005) Strategies for mapping and cloning quantitative trait genes in rodents. Nature reviews. Genetics, 6, 271-286.

41. Stadnicka, A., Contney, S.J., Moreno, C., Weihrauch, D., Bosnjak, Z.J., Roman, R.J. and Stekiel, T.A. (2009) Mechanism of differential cardiovascular response to propofol in Dahl salt-sensitive, Brown Norway, and chromosome 13-substituted consomic rat strains: role of large conductance Ca2+ and voltage-activated potassium channels. The Journal of pharmacology and experimental therapeutics, 330, 727-735.

42. Rosengren, A.H., Jokubka, R., Tojjar, D., Granhall, C., Hansson, O., Li, D.Q., Nagaraj, V., Reinbothe, T.M., Tuncel, J., Eliasson, L. et al. (2010) Overexpression of alpha2A-adrenergic receptors contributes to type 2 diabetes. Science, 327, 217-220.

43. The Rat Genome Database. http://rgd.mcw.edu/.

44. The Rat Genome Project. https://www.hgsc.bcm.edu/other-mammals/rat-genome-project.

45. van Heesch, S., Kloosterman, W.P., Lansu, N., Ruzius, F.P., Levandowsky, E., Lee, C.C., Zhou, S., Goldstein, S., Schwartz, D.C., Harkins, T.T. et al. (2013) Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. BMC genomics, 14, 257.

46. Shendure, J. and Lieberman Aiden, E. (2012) The expanding scope of DNA sequencing. Nature biotechnology, 30, 1084-1094.

47. Atanur, S.S., Birol, I., Guryev, V., Hirst, M., Hummel, O., Morrissey, C., Behmoaras, J., Fernandez-Suarez, X.M., Johnson, M.D., McLaren, W.M. et al. (2010) The genome sequence of the spontaneously hypertensive rat: Analysis and functional significance. Genome research, 20, 791-803.

48. Simonis, M., Atanur, S.S., Linsen, S., Guryev, V., Ruzius, F.P., Game, L., Lansu, N., de Bruijn, E., van Heesch, S., Jones, S.J. et al. (2012) Genetic basis of transcriptome differences between the founder strains of the rat HXB/BXH recombinant inbred panel. Genome biology, 13, r31.

49. Atanur, S.S., Diaz, A.G., Maratou, K., Sarkis, A., Rotival, M., Game, L., Tschannen, M.R., Kaisaki, P.J., Otto, G.W., Ma, M.C. et al. (2013) Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. Cell, 154, 691-703.

50. Guo, X., Brenner, M., Zhang, X., Laragione, T., Tai, S., Li, Y., Bu, J., Yin, Y., Shah, A.A., Kwan, K. et al. (2013) Whole-genome sequences of DA and F344 rats with different susceptibilities to arthritis, autoimmunity, inflammation and cancer. Genetics, 194, 1017-1028.

51. Backdahl, L., Ekman, D., Jagodic, M., Olsson, T. and Holmdahl, R. (2014) Identification of candidate risk gene variations by whole-genome sequence analysis of four rat strains commonly used in inflammation research. BMC genomics, 15, 391.

52. Ma, M.C., Atanur, S.S., Aitman, T.J. and Kwitek, A.E. (2014) Genomic structure of nucleotide diversity among Lyon rat models of metabolic syndrome. BMC genomics, 15, 197.

53. Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M. et al. (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature, 477, 289-294.

54. Nellaker, C., Keane, T.M., Yalcin, B., Wong, K., Agam, A., Belgard, T.G., Flint, J., Adams, D.J., Frankel, W.N. and Ponting, C.P. (2012) The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome biology, 13, R45.

55. Yalcin, B., Wong, K., Agam, A., Goodson, M., Keane, T.M., Gan, X., Nellaker, C., Goodstadt, L., Nicod, J., Bhomra, A. et al. (2011) Sequence-based characterization of structural variation in the mouse genome. Nature, 477, 326-329.

56. Yalcin, B., Wong, K., Bhomra, A., Goodson, M., Keane, T.M., Adams, D.J. and Flint, J. (2012) The fine-scale architecture of structural variants in 17 mouse genomes. Genome biology, 13, R18.

57. Su, Z., Li, Z., Chen, T., Li, Q.Z., Fang, H., Ding, D., Ge, W., Ning, B., Hong, H., Perkins, R.G. et al. (2011) Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. Chemical research in toxicology, 24, 1486-1493.

58. Yu, Y., Fuscoe, J.C., Zhao, C., Guo, C., Jia, M., Qing, T., Bannon, D.I., Lancashire, L., Bao, W., Du, T. et al. (2014) A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. Nature communications, 5, 3230.

59. Heinz, S., Romanoski, C.E., Benner, C., Allison, K.A., Kaikkonen, M.U., Orozco, L.D. and Glass, C.K. (2013) Effect of natural genetic variation on enhancer selection and function. Nature, 503, 487-492.

60. Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V. et al. (2013) Extensive variation in chromatin states across humans. Science, 342, 750-752.

61. Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I. et al. (2013) Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science, 342, 744-747.

62. McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y. and Pritchard, J.K. (2013) Identification of genetic variants that affect histone modifications in human cells. Science, 342, 747-749.

63. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57-74.

64. Kutter, C., Brown, G.D., Goncalves, A., Wilson, M.D., Watt, S., Brazma, A., White, R.J. and Odom, D.T. (2011) Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. Nature genetics, 43, 948-955.

65. Hull, R.P., Srivastava, P.K., D'Souza, Z., Atanur, S.S., Mechta-Grigoriou, F., Game, L., Petretto, E., Cook, H.T., Aitman, T.J. and Behmoaras, J. (2013) Combined ChIP-Seq and transcriptome analysis identifies AP-1/JunD as a primary regulator of oxidative stress and IL-1beta synthesis in macrophages. BMC genomics, 14, 92.

66. Rintisch, C., Heinig, M., Bauerfeind, A., Schafer, S., Mieth, C., Patone, G., Hummel, O., Chen, W., Cook, S., Cuppen, E. et al. (2014) Natural variation of histone modification and its impact on gene expression in the rat genome. Genome research, 24, 942-953.

67. Srivastava, P.K., Hull, R.P., Behmoaras, J., Petretto, E. and Aitman, T.J. (2013) JunD/AP1 regulatory network analysis during macrophage activation in a rat model of crescentic glomerulonephritis. BMC systems biology, 7, 93.

68. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485, 376-380.

69. Kitada, K., Ishishita, S., Tosaka, K., Takahashi, R., Ueda, M., Keng, V.W., Horie, K. and Takeda, J. (2007) Transposon-tagged mutagenesis in the rat. Nature methods, 4, 131-133.

70. Smits, B.M., Mudde, J.B., van de Belt, J., Verheul, M., Olivier, J., Homberg, J., Guryev, V., Cools, A.R., Ellenbroek, B.A., Plasterk, R.H. et al. (2006) Generation of gene knockouts and mutant models in the laboratory rat by ENU-driven target-selected mutagenesis. Pharmacogenetics and genomics, 16, 159-169.

71. Mashimo, T., Yanagihara, K., Tokuda, S., Voigt, B., Takizawa, A., Nakajima, R., Kato, M., Hirabayashi, M., Kuramoto, T. and Serikawa, T. (2008) An ENU-induced mutant archive for gene targeting in rats. Nature genetics, 40, 514-515.

72. Capecchi, M.R. (2005) Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. Nature reviews. Genetics, 6, 507-512.

73. Buehr, M., Meek, S., Blair, K., Yang, J., Ure, J., Silva, J., McLay, R., Hall, J., Ying, Q.L. and Smith, A. (2008) Capture of authentic embryonic stem cells from rat blastocysts. Cell, 135, 1287-1298.

74. Li, P., Tong, C., Mehrian-Shai, R., Jia, L., Wu, N., Yan, Y., Maxson, R.E., Schulze, E.N., Song, H., Hsieh, C.L. et al. (2008) Germline competent embryonic stem cells derived from rat blastocysts. Cell, 135, 1299-1310.

75. Li, W., Wei, W., Zhu, S., Zhu, J., Shi, Y., Lin, T., Hao, E., Hayek, A., Deng, H. and Ding, S. (2009) Generation of rat and human induced pluripotent stem cells by combining genetic reprogramming and chemical inhibitors. Cell stem cell, 4, 16-19.

76. Liao, J., Cui, C., Chen, S., Ren, J., Chen, J., Gao, Y., Li, H., Jia, N., Cheng, L., Xiao, H. et al. (2009) Generation of induced pluripotent stem cell lines from adult rat cells. Cell stem cell, 4, 11-15.

77. Tong, C., Li, P., Wu, N.L., Yan, Y. and Ying, Q.L. (2010) Production of p53 gene knockout rats by homologous recombination in embryonic stem cells. Nature, 467, 211-213.

78. Tong, C., Huang, G., Ashton, C., Li, P. and Ying, Q.L. (2011) Generating gene knockout rats by homologous recombination in embryonic stem cells. Nature protocols, 6, 827-844.

79. Geurts, A.M., Cost, G.J., Freyvert, Y., Zeitler, B., Miller, J.C., Choi, V.M., Jenkins, S.S., Wood, A., Cui, X., Meng, X. et al. (2009) Knockout rats via embryo microinjection of zinc-finger nucleases. Science, 325, 433.

80. Tesson, L., Usal, C., Menoret, S., Leung, E., Niles, B.J., Remy, S., Santiago, Y., Vincent, A.I., Meng, X., Zhang, L. et al. (2011) Knockout rats generated by embryo microinjection of TALENs. Nature biotechnology, 29, 695-696.

81. Li, D., Qiu, Z., Shao, Y., Chen, Y., Guan, Y., Liu, M., Li, Y., Gao, N., Wang, L., Lu, X. et al. (2013) Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. Nature biotechnology, 31, 681-683.

82. Yoshimi, K., Kaneko, T., Voigt, B. and Mashimo, T. (2014) Allele-specific genome editing and correction of disease-associated phenotypes in rats using the CRISPR-Cas platform. Nature communications, 5, 4240.

83. Brown, A.J., Fisher, D.A., Kouranova, E., McCoy, A., Forbes, K., Wu, Y., Henry, R., Ji, D., Chambers, A., Warren, J. et al. (2013) Whole-rat conditional gene knockout via genome editing. Nature methods, 10, 638-640.

84. EURATRANS: European large-scale functional genomics in the rat for translational research. http://www.euratrans.eu/.

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

86. Zhou, Y., Xu, J., Liu, Y., Li, J., Chang, C. and Xu, C. (2014) Rat hepatocytes weighted gene co-expression network analysis identifies specific modules and hub genes related to liver regeneration after partial hepatectomy. PloS one, 9, e94868.

2

Chapter 2

Genomic landscape of rat strain and substrain variation

Roel Hermsen[1], Joep de Ligt[1], Wim Spee[1], Francis Blokzijl[1], Sebastian Schäfer[2], Eleonora Adami[2], Sander Boymans[1], Stephen Flink[3], Ruben van Boxtel[1], Robin H. van der Weide[1], Tim Aitman[4], Norbert Hübner[2], Marieke Simonis[1], Boris Tabakoff[3], Victor Guryev[5], Edwin Cuppen[1]

1 Hubrecht Institute, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands.

2 Max Delbrück Center for Molecular Medicine, Berlin, Germany

3 University of Colorado School of Medicine, Department of Pharmacology, 12800 E. 19th Ave. Aurora, CO, USA.

4 Physiological Genomic and Medicine Group, MRC Clinical Sciences Centre

5 European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen, Antonius Deusinglaan 1, 9713 AD Groningen, The Netherlands.

Abstract

Background

Since the completion of the rat reference genome in 2003, whole-genome sequencing data from more than 40 rat strains have become available. These data represent the broad range of strains that are used in rat research including commonly used substrains. Currently, this wealth of information cannot be used to its full extent, because the variety of different variant calling algorithms employed by different groups impairs comparison between strains. In addition, all rat whole genome sequencing studies to date have used an outdated reference genome for analysis (RGSC3.4 released in 2004).

Results

Here we present a comprehensive, multi-sample and uniformly called set of genetic variants in 40 rat strains, including 19 substrains. We reanalyzed all primary data using a recent version of the rat reference assembly (RGSC5.0 released in 2012) and identified over 12 million genomic variants (SNVs, indels and structural variants) among the 40 strains. 28,318 SNVs are specific to individual substrains, which may be explained by introgression from other unsequenced strains and ongoing evolution by genetic drift. Substrain SNVs tend to have a larger predicted functional impact compared to older shared SNVs.

Conclusions

In summary we present a comprehensive catalog of uniformly analyzed genetic variants among 40 widely used rat inbred strains based on the RGSC5.0 assembly. This represents a valuable resource, which will facilitate rat functional genomic research. In line with previous observations, our genome-wide analyses do not show evidence for contribution of multiple ancestral founder rat subspecies to the currently used rat inbred strains, as is the case for mouse. In addition, we find that the degree of substrain variation is highly variable between strains, which is of importance for the correct interpretation of experimental data from different labs.

Introduction

The rat is an important model organism for studying human disease biology [1]. In the past century, a great variety of strains and substrains have been bred that differ in susceptibility to complex diseases like hypertension, diabetes, autoimmunity, cancer and addiction disorders. Due to practical limitations, studies on disease phenotypes are often conducted in varying substrains by different research groups. For example, SHR/NCrl and SHR/NHsd are used for studying cardiovascular phenotypes in the United States [2] and Europe [3], respectively. The effect on the interpretability and extrapolation of the obtained results by using different substrains remains unclear. Several studies based on DNA SNP marker panels showed that genetic variation between substrains is present [4-6]. However, the magnitude of this difference can only be properly interpreted when assessed on a genome-wide scale as variation is not necessarily randomly distributed throughout the genome. Here, we systematically (re)analyzed whole genome sequence (WGS) data of 40 rat strains and substrains resulting in a comprehensive inventory of strain and substrain-specific variants.

With the emergence of next-generation sequencing (NGS) techniques many rat strains and substrains have been whole genome sequenced [7-12], with the primary goal to provide insight in the genetic factors underlying phenotypic traits in these strains. After the availability of the first rat reference genome assembly in 2003 [13], the first variation catalog of a non-reference inbred strain, the spontaneously hypertensive rat (SHR), was published in 2010 [7]. This data was later combined with the BN-*Lx* genome sequence and extended with RNA sequencing data, resulting in a comprehensive catalog of genetic variation and associated quantitative and qualitative transcription phenotypes, in the HXB/BXH recombinant inbred (RI) panel [8]. This panel is a valuable tool for dissection of the complex genetic basis of cardiovascular, behavioral, and developmental disorders. In addition, the eight founders of the rat heterogeneous stock (NIH-HS) were recently sequenced [9]. In this study, the genome sequence of the founder strains were used to impute the genomes of the 1407 SNP-genotyped heterogeneous stock rats that were also extensively phenotyped. This work resulted in the identification of 355 high-resolution quantitative trait loci (QTLs) for 122 phenotypes.

More rat whole genome sequence data became available by publication of the variation catalog and strain specific sequences of the Dark Agouti (DA) and Fischer (F344) rat, which carry unique dichotomous phenotypes, such as rheumatoid arthritis and several cancer types [10]. Finally, a large community-driven effort in rat genome sequencing yielded variation catalogs of 25 inbred strains and substrains [11]. Analysis of this data identified strain-specific selective sweeps and gene clusters that implied genes involved in the development of cardiovascular disease in rat.

One important factor that determines the success of cataloging genomic variation is the quality of the used reference genome. Since its initial publication in 2003, the rat reference genome has undergone major improvements and was recently further improved using a range of NGS-based methods [14]. This has resulted in version 5.0 of the rat reference assembly in 2012 [15]. Although the v5.0 assembly shows great overall improvement at both nucleotide and the structural level, it has not yet been used as a reference for the analysis of the aforementioned rat genomes. Instead, these studies all used the v3.4 assembly, which is publicly available since 2004 [13] and contains many gaps, assembly inconsistencies and nucleotide and indel errors (due to the relatively low coverage and typical errors associated with capillary dideoxy sequencing).

Finally, bioinformatic analysis of whole genome sequencing data, including mapping and variant calling, has matured rapidly over the past years. However, as a result of these ongoing developments, a broad range of bioinformatic tools and settings were used for the analysis of currently published rat genomes. Direct comparison of different strains therefore becomes challenging, especially because many old tools did not call reference positions. Taken together, a comprehensive overview and systematic comparison of laboratory rat genomic variation is currently lacking. Such a resource would be useful for a broad range of rat researchers, as it allows proper selection of experimental and control rat strains and interpretation of potential substrain effects in published experiments.

Results

Genetic variation among strains

We collected the genomes of 37 rat strains that were sequenced previously [7-12] (Table 1) and analyzed them together with newly derived sequences from the BN-*Lx*/CubPrin, SHR/OlaIpcvPrin and SHR/NCrlPrin rat strains (Supplemental table 1). We aligned reads of all 40 strains to the RGSC5.0 assembly (BN/NHsdMcWi; [13]). After applying strict criteria (see Methods) and using multi-sample variant calling we identified in total 9,183,702 SNVs, 3,001,935 indels and 63,664 structural variants compared to the reference assembly. To assess the sensitivity and specificity of our calls we made use of finished capillary sequencing data from 13 BAC clones from the LE/Stm strain, which was also sequenced by two different NGS approaches. We evaluated 2,132,438 nucleotides and found in total 2,468 SNVs that were detected by capillary sequencing and NGS techniques. 141 SNVs were missed by whole-genome sequencing; resulting in an estimate of 524,677 (5.4%) missed SNVs genome-wide. 14 SNVs identified by whole-genome sequencing were not found in the BACs; resulting in an estimate of 55,817 (0.6%) false positive SNV calls genome-wide. For indels the false positive and negative call rates are higher (FP:15,7% FN:27,3%) due to known detection difficulties of current calling algorithms. Although the 40 strains were sequenced on two different NGS platforms (SOLiD and Illumina), false positive and negative call rates based on the LE data (sequenced on both platforms) were similar (Supplemental table 2).

Small genomic variation: SNVs and indels

We identified single nucleotide variants and small insertions and deletions (indels) with the Genome Analysis Toolkit (GATK) HaplotypeCaller [16]. All together we identified 9.2M SNVs of which 97.5% were homozygous and 2.5% were heterozygous. This small percentage of heterozygous variants can be attributed to incomplete fixation of the inbred strain, genomic duplications followed by diversification, and technical errors in the sequencing or data analysis. These variants were filtered out in a separate file (see Availability of Supporting Data) and were not taken into account in further downstream analyses.

Rat strain	Publication	PMID	Sequencing platform	Number of SNVs	Number of indels	Number of Structural Variants
ACI/EurMcwi	Atanur et al	23890820	Illumina HiSeq2X00	3,539,775	1,651,251	7,259
ACI/N	Baud et al	23708188	SOLiD 4 and 5500	3,125,523	1,382,793	19,541
BBDP/Wor	Atanur et al	23890820	Illumina HiSeq2X00	3,279,444	1,526,223	3,678
BN/SsN	Baud et al	23708188	SOLiD 4 and 5500	59,402	660,918	14,126
BN-*Lx*/Cub	Simonis et al; Atanur et al	22541052; 23890820	SOLiD 2,3 and 4	102,359	627,056	13,391
BN-*Lx*/ CubPrin	Hermsen et al	na	Illumina HiSeq2000	140,376	420,433	13,410
BUF/N	Baud et al	23708188	SOLiD 4 and 5500	2,848,992	1,302,710	18,481
DA/BklArbNsi	Guo et al	23695301	Illumina HiSeq2000	3,368,008	1,567,160	4,184
F334/N	Baud et al	23708188	SOLiD 4 and 5500	2,947,509	1,342,709	20,881
F344/NCrl	Atanur et al	23890820	Illumina HiSeq2X00	3,369,205	1,579,418	3,492
F344/NHsd	Guo et al	23695301	Illumina HiSeq2000	3,367,166	1,573,573	3,950
FHH/EurMcwi	Atanur et al	23890820	Illumina HiSeq2X00	3,389,304	1,592,915	3,011
FHL/EurMcwi	Atanur et al	23890820	Illumina HiSeq2X00	3,361,824	1,586,543	8,504
GK/Ox	Atanur et al	23890820	Illumina HiSeq2X00	3,549,952	1,575,619	4,241
LE/Stm (Illumina)	Atanur et al	23890820	Illumina HiSeq2X00	3,412,610	1,578,099	2,598
LE/Stm (SOLiD)	Baud et al	23708188	SOLiD 4 and 5500	2,949,814	1,359,947	21,038
LEW/Crl	Atanur et al	23890820	Illumina HiSeq2X00	2,884,477	1,409,659	3,642
LEW/NCrlBR	Atanur et al	23890820	Illumina HiSeq2X00	2,884,763	1,402,459	3,996
LH/MavRrrc	Atanur et al; Ma et al	23890820; 24628878	Illumina HiSeq2X00	3,369,852	1,584,236	2,891
LL/MavRrrc	Atanur et al; Ma et al	23890820; 24628878	Illumina HiSeq2X00	3,329,343	1,565,343	3,070
LN/MavRrrc	Atanur et al; Ma et al	23890820; 24628878	Illumina HiSeq2X00	3,319,381	1,562,698	2,952
M520/N	Baud et al	23708188	SOLiD 4 and 5500	2,896,825	1,321,431	19,308

Rat strain	Publication	PMID	Sequencing platform	Number of SNVs	Number of indels	Number of Structural Variants
MHS/Gib	Atanur et al	23890820	Illumina HiSeq2X00	3,183,312	1,513,330	2,917
MNS/Gib	Atanur et al	23890820	Illumina HiSeq2X00	3,168,796	1,538,413	3,105
MR/N	Baud et al	23708188	SOLiD 4 and 5500	2,878,806	1,350,411	18,001
SBH/Ygl	Atanur et al	23890820	Illumina HiSeq2X00	3,393,610	1,617,252	14,787
SBN/Ygl	Atanur et al	23890820	Illumina HiSeq2X00	3,300,171	1,592,247	15,216
SHR/NCrlPrin	Hermsen et al	na	Illumina HiSeq2000	3,736,435	1,694,012	14,179
SHR/NHsd	Atanur et al	23890820	Illumina HiSeq2X00	3,756,155	1,705,126	3,950
SHR/OlaIpcv	Simonis et al; Atanur et al	22541052; 23890820	Illumina Genome Analyser 2	3,747,579	1,706,963	4,066
SHR/ OlaIpcvPrin	Hermsen et al	na	Illumina HiSeq2000	3,709,362	1,689,758	14,069
SHRSP/Gla	Atanur et al	23890820	Illumina HiSeq2X00	3,700,495	1,723,961	2,301
SR/Jr	Atanur et al	23890820	Illumina HiSeq2X00	3,353,579	1,568,778	3,699
SS/Jr	Atanur et al	23890820	Illumina HiSeq2X00	3,311,117	1,553,050	3,685
SS/ JrHsdMcwi	Atanur et al	23890820	Illumina HiSeq2X00	3,310,209	1,595,799	7,938
SUO_F344	Hermsen et al	na	Illumina HiSeq2000	3,349,024	1,549,272	11,864
WAG/Rij	Atanur et al	23890820	Illumina HiSeq2X00	3,092,505	1,485,673	3,650
WKY/Gla	Atanur et al	23890820	Illumina HiSeq2X00	3,777,400	1,725,868	3,292
WKY/N	Baud et al	23708188	SOLiD 4 and 5500	3,213,913	1,419,460	21,832
WKY/NCrl	Atanur et al	23890820	Illumina HiSeq2X00	3,502,459	1,700,646	3,630
WKY/NHsd	Atanur et al	23890820	Illumina HiSeq2X00	3,682,736	1,665,949	4,691
WN/N	Baud et al	23708188	SOLiD 4 and 5500	2,899,096	1,323,116	18,995

Table 1. Summary of the genomic variation of the included rat strains in this study.

Type	Impact	Count	Fraction	Sum
Stop gained	High	285	0.0%	696
Splice site donor		209	0.0%	
Splice site acceptor		158	0.0%	
Start lost		26	0.0%	
Stop lost		18	0.0%	
Non synonymous coding	Moderate	26,239	0.3%	26,239
Synonymous coding	Low	42,182	0.4%	42,947
Start gained		725	0.0%	
Synonymous stop		35	0.0%	
Non synonymous start		5	0.0%	
Intergenic	Modifier	6,509,332	62.2%	10,394,771
Intron		2,991,180	28.6%	
Downstream		430,875	4.1%	
Upstream		427,613	4.1%	
UTR 3 Prime		27,145	0.3%	
UTR 5 Prime		4,357	0.0%	
Exon		4,269	0.0%	
Total effects		*10,464,653*		*10,464,653*

Table 2. Summary of the effect prediction of the detected SNVs.

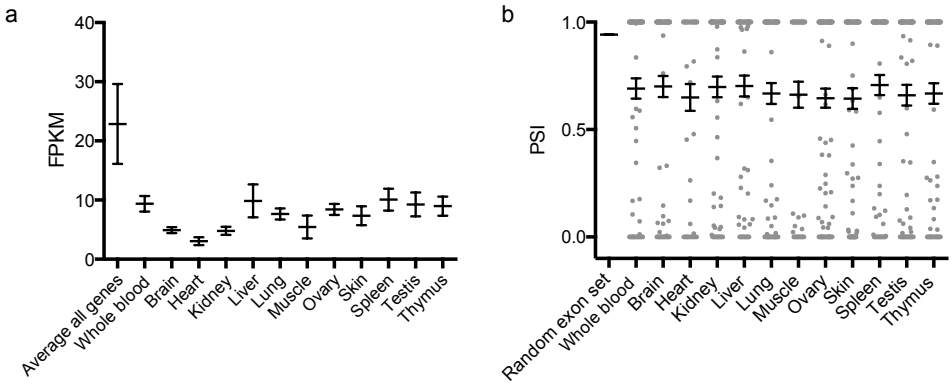


Figure 1 – 'Repression' of genes and exons containing high impact SNVs. (a) Genome-wide average FPKM ± SEM across all tissues compared to the average FPKM of genes containing high impact SNVs for 12 tissues. Genes containing high impact SNVs are significantly lower expressed (Non-parametric ANOVA; $p<0.0001$). (b) The average Percentage Spliced In (PSI) ± SEM across the transcriptome was compared to the average PSI of exons containing high impact SNVs for 12 tissues. Exons containing high impact SNVs are significantly more spliced out/not used (Non-parametric ANOVA; $P < 0.0001$).

To understand the functional consequences of the SNVs we annotated these variants using SnpEff (Table 2) [17]. Predictions on the functional consequences of a variant are typically overestimated due to for instance their presence in pseudogenes or non-constitutive exons [18]. Here we set out to systematically characterize 601 SNVs which are annotated to have a deleterious effect (marked as causing "HIGH" impact by SnpEff) on gene function including stop-gain mutations and alterations of splice sites (Table 2). First we tested the hypothesis that neighboring variants could possibly restore the open reading frame by investigating the high impact SNV vicinity. We found for 60 SNVs (10%) a neighboring SNV or indel that restored the open reading frame (Supplemental table 3). From the remaining 541 high impact SNVs we determined the expression in twelve BN-*Lx*/Cub tissues for the genes in which the variants are located (Figure 1). We then compared this to the expression of all genes and found that the highly impacted genes are expressed at significantly lower levels (non-parametric ANOVA; $p<0.0001$). In addition, for the expressed genes, we analyzed the usage of individual exons by means of the ''Percentage Spliced In' (PSI) index per exon. Interestingly, we found that the exons containing high impact SNVs tend to be less used and more often spliced out than expected (non-parametric ANOVA; $P < 0.0001$). Thus, we conclude that most high impact SNVs will actually only have a limited biological relevance, in part by neutralization by neighboring variants or by being 'repressed' in expression at the gene and exon levels.

Cross-species comparison of genome variation

To get an impression of the nucleotide diversity among laboratory rat strains in relation to other domesticated animals, we compared the SNV density between five different domesticated species. We extracted all autosomal genomic regions that are one-to-one comparable (syntenic) with the rat genome from dog, horse, pig and mouse. Next, we determined the amount of species-specific SNVs in each 100 kilobase syntenic window to identify regions that contain high and low nucleotide diversity in each species. We extracted the regions with highest and lowest amount of SNVs that are shared among all five species. In total, the cumulative regions with a low SNV density contain 22 genes at 4 genomic loci (Figure 2a). When we functionally annotate these genes using the DAVID algorithm [19], we find enrichment ($p<0.001$) for genes involved in catabolic processes (Supplemental table 4). This might reflect the evolutionary constraint on diet, exerted in these five species by domestication [20]. For the regions

that exhibit high SNV density in all five species we in total find 51 genes at 6 genomic loci (Figure 2b). Functional annotation with DAVID shows an enrichment ($p<0.001$) for olfactory and hemoglobin genes, which are known to rapidly evolve and are highly variable in several species [21, 22]. Another way to look at loci under selective pressure is by studying the non-synonymous to synonymous substitution rate per gene (K_a/K_s ratio). Genes that are potentially under positive selection have a non-synonymous to synonymous ratio of >1.0 [23]. We identified all protein coding genes (n=22,941) that contain 6 or more SNVs in the protein-coding region (n=3,006) and extracted the genes that have a non-synonymous to synonymous ratio of >1.0 (n=909). DAVID functional annotation clustering of these 909 genes using the 3,006 genes as background shows that this set is enriched for genes related to the olfactory system ($p<0.001$; Supplemental table 5). This data confirms the results of the interspecies SNV density analysis and shows that within rat strains these types of genes are indeed highly polymorphic [21].

'Population' structure

To get an impression of the 'population' structure of these 40 strains, we used the SNV genotype information per locus in a Bayesian approach to define clusters without any other prior knowledge. In addition, to demonstrate the power of this approach to accurately define clusters, we included genotypes from WGS data from a Strain of Unknown Origin (SUO). We hypothesized that we would be able to designate the strain of origin based on the genotypes of a broad representation of rat strains in this data set. We performed this analysis using fastStructure, which is an algorithm for inferring population structure from large SNP genotype data [24]. fastStructure identifies the number of populations (clusters or 'K') needed to explain the structure in the data in which individual samples can have membership in multiple clusters. When we analyze the genotypes of all 40+1 rat strains we find that we can differentiate nine distinct clusters (Figure 3a). Five strains have membership in multiple clusters, which may reflect shared ancestry or interbreeding before or during inbred strain derivation, whereas the other strains only consist of one cluster. In general most clusters resemble the previously published classification based on a rooted phylogenetic tree [11]. In addition this method allows identification of similarity between clusters that have been separated in a phylogenetic tree analysis.

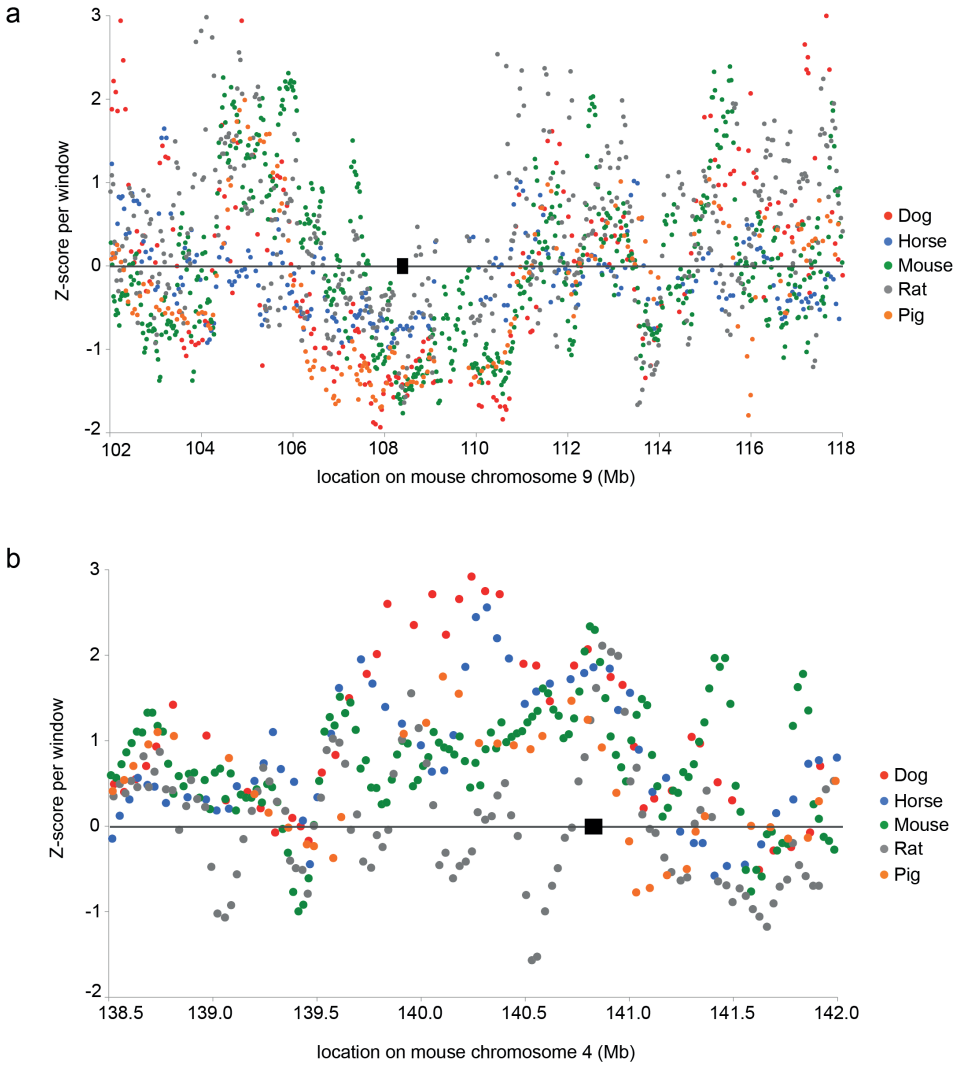


Figure 2 – Cross-species comparison of SNV densities. (a) An example of a locus (black rectangle) on mouse chromosome 9 with the lowest SNV density in five domesticated species. (b) An example of a locus (black rectangle) on mouse chromosome 4 with the highest SNV density in five domesticated species.

For example, the GK/Ox strain, which is a Wistar derived strain originating from Japan, also shows contribution of the cluster which contains the Wistar derived strains from Europe and the United States [11]. We also find that the included SUO strain clearly shows a full match in the Fischer (F344) cluster and we therefore conclude that the SUO is a substrain of the F344 strain (SUO_F344). Besides the ancestral clustering of strains, we also studied the subchromosomal pattern of similarity and divergence.

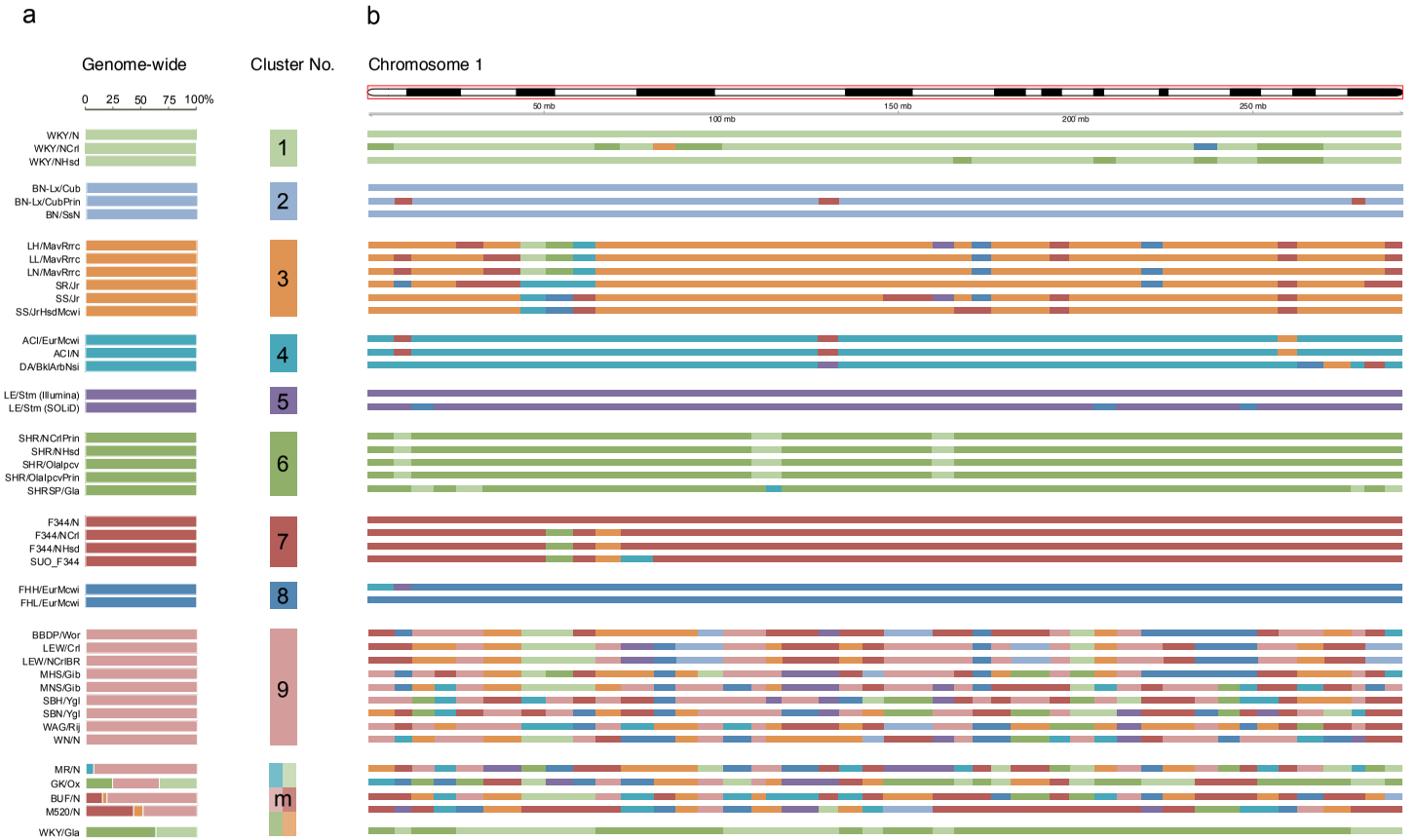

a
b
Genome-wide
Cluster No.
Chromosome 1
0 25 50 75 100%
50 mb
100 mb
150 mb
200 mb
250 mb
WKY/N
WKY/NCrl
WKY/NHsd
BN-Lx/Cub
BN-Lx/CubPrin
BN/SsN
LH/MavRrrc
LL/MavRrrc
LN/MavRrrc
SR/Ur
SS/Ur
SS/UrHsdMcwi
ACI/EurMcwi
ACI/N
DA/BklArbNsi
LE/Stm (Illumina)
LE/Stm (SOLiD)
SHR/NCrlPrin
SHR/NHsd
SHR/OlaIpcv
SHR/OlaIpcvPrin
SHRSP/Gla
F344/N
F344/NCrl
F344/NHsd
SUO_F344
FHH/EurMcwi
FHL/EurMcwi
BBDP/Wor
LEW/Crl
LEW/NCrlBR
MHS/Gib
MNS/Gib
SBH/Ygl
SBN/Ygl
WAG/Rij
WN/N
MR/N
GK/Ox
BUF/N
M520/N
WKY/Gla
1
2
3
4
5
6
7
8
9
m

Figure 3 – 'Population' structure of 41 rat strains. (a) Per strain, the contribution from the 9 different clusters is plotted as percentage of the genome. Each cluster is represented by a separate color. The cluster designated with a 'm' represents the strains that have membership from multiple clusters. (b) Per strain, the genomic distribution along rat chromosome 1 is plotted as an example. The colors match the cluster colors from (a).

We determined for each bin of 20,000 SNVs to which cluster it was most similar (Figure 3b and Supplemental figure 1). Based on this analysis we observed that the overall clustering based on the genomes as a whole, matches the clusters found in the genomic cluster distribution using the 40+1 strains and is concordant with previous work [11]. We find that substrains (e.g. the SHR substrains) have a comparable genomic cluster structure, indicating recent divergence. Of note, the relatively large window size of 20,000 SNVs may cause overrepresentation of differential loci between substrains that are known to be very similar (e.g. the Lyon strains [12]). Nevertheless, we find five rat strains that showed contribution from multiple clusters in the fastStructure analysis (group 'm') of which one (WKY/Gla) shows a genomic distribution of the clusters #1 (WKY) and #6 (SHR), which is in line with its known breeding origin [11, 25]. In addition, cluster 9 (with e.g. the LEW substrains) shows a confetti-like signature, while the fastStructure analysis does not categorize them as multi-cluster strains. In conclusion, we see shared haplotypes between strains in different clusters, indicating common ancestry and/or cross-breeding during inbred strain derivation. Nevertheless, the variation uniqueness per cluster is very high.

Large genomic variation: structural variants

Structural variants were determined using two independent methods. I) We interrogated the orientation of the mapped read-pairs (RP) compared to the reference genome to detect deletions, tandem duplications and inversions by applying the DELLY [26] algorithm in all samples simultaneously. II) CNVnator [27] was used to identify relative changes in read-depth (RD) thereby detecting losses and gains of genomic segements. Given the algorithmic difficulties in detecting structural variants we took a strict cutoff to minimize false positive calls (see Methods). In total, we identified 34,433 deletions, 585 tandem duplications and 26,899 inversions based on the read-pair method together with 1,747 copy number variable sites based on the read-depth method. All together this resulted in 63,664 SVs in the 40 strains.

Substrain variability

To identify the genomic variants that differ between substrains we used the seven strains of which data for at least two substrains was available: ACI, BN, F344, LEW, SHR, SS and WKY (Table 3). We did not include WKY/Gla because this substrain is known to have diverged significantly from the other WKY substrains [25] which is also evident from our genomic comparisons. For each of the seven groups we identified all positions that were variable between the substrains. We found that the degree of substrain variation was highly variable between strains (1,046-10,250 per strain) (Table 3), which may reflect the time after separation of the substrain colonies. For comparative functional analyses of substrain variation (detailed below) we used all other SNVs (8,863,815), excluding variants that were shared by all strains, as a comparison group.

Strain	Substrains	Substrain SNVs
ACI	ACI/N	3,432
	ACI/EurMcwi	
BN	BN-*Lx*/Cub	2,291
	BN-*Lx*/CubPrin	
	BN/SsN	
F344	F334/N	5,854
	F344/NHsd	
	F344/NCrl	
	SUO_F344	
LEW	LEW/Crl	1,046
	LEW/NCrlBR	
SHR	SHR/OlaIpcv	2,950
	SHR/NCrlPrin	
	SHR/NHsd	
	SHR/OlaIpcvPrin	
SS	SS/Jr	2,495
	SS/JrHsdMcwi	
WKY	WKY/N	10,250
	WKY/NCrl	
	WKY/NHsd	
Total		*28,318*

Table 3. Summary of substrain SNVs

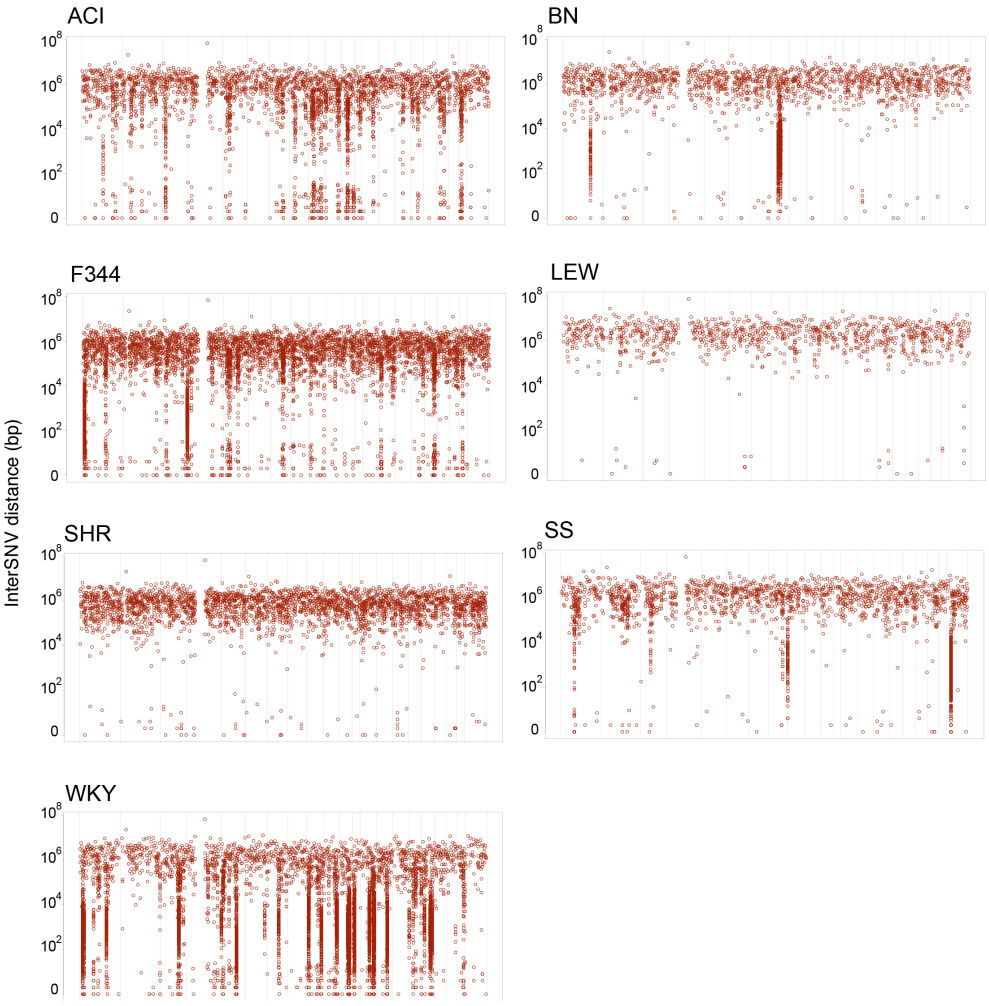


Figure 4 – Genomic distribution of substrain variants per strain. For each strain the distance between two consecutive SNVs (y-axis) is plotted along the genomic position (x-axis). The windows on the x-axis represent the different chromosomes. Loci with a high density of substrain SNVs can be observed as clusters that drop down from the average genome-wide pattern.

To get an impression of the genomic distribution of the substrain SNVs we plotted the genomic distance between two consecutive substrain SNVs (Figure 4). For two groups (LEW and SHR) we found an even distribution of the SNVs through the genome, while in the other five groups we also observe clustering of SNVs. This effect is limited to a few loci for BN, but is more widespread for WKY. One explanation for the clustering of these SNVs can be introgression from a rat strain that is not included in the current analysis. For instance we observe a cluster of SNVs in the BN

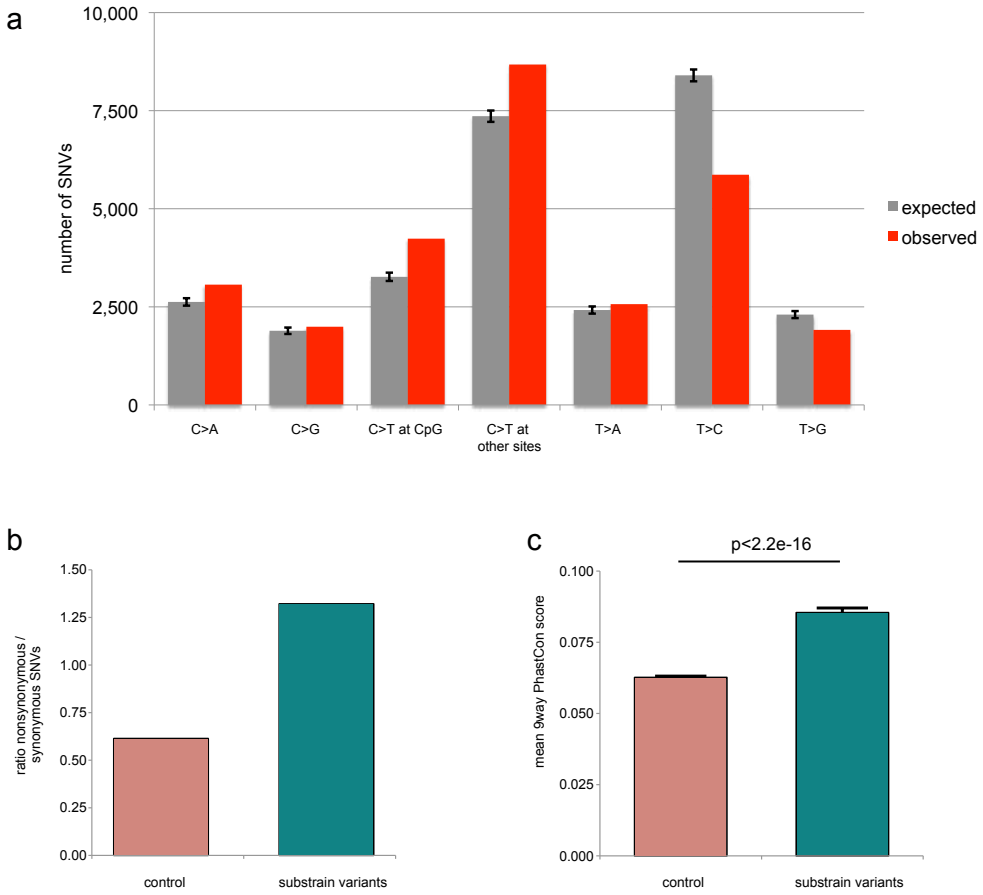


Figure 5 – Substrain variant characteristics. (a) Bar plots showing the contribution of each nucleotide change for all substrain variants (observed) versus the control variants (expected). Error bars represent the 95% confidence interval. (b) Bar plot showing the K_a/K_s ratio ratio of the substrain variants versus the control variants. (c) Bar plot showing the average phastCons score for each substrain variant compared to the control variants. Substrain variants affect nucleotides with a significantly higher phastCons score (Student's t-test; p<2.2e-16). Error bars represent the SEM.

group on chromosome 8. For the BN-*Lx* substrains that are in this group, this region is known to contain the *Lx* locus from the polydactylous PD/Cub strain [28]. Since whole genome sequencing data of the PD/Cub strain is not available we observe the congenic *Lx* segment as an introgressed cluster of substrain-specific SNVs in our analysis. Although this analysis is able to identify introgressed loci from other sequenced strains, we cannot exclude that we miss introgression from closely related strains with limited SNV diversity.

Besides introgression, the occurrence of *de novo* mutations (genetic drift) appears the main driver of substrain variation [29]. To understand the process of newly arising variants we analyzed the different types of nucleotide changes that occured. The control set of 8,863,815 SNVs was used to estimate the expected amount of substitutions per category. The observed amount of nucleotide changes of the 28,318 substrain SNVs was then compared to this expected pattern. We find an enrichment of C to T substitutions in general, which is most pronounced at CpG dinucleotides (Figure 5a). This may reflect an elevated rate of spontaneous/oxidative deamination of 5-methyl-cytosines, which is associated with oxidative DNA damage in animal genomes [30]. In addition, we find a significant depletion of T to C changes, which are typically the result of alkylating mechanisms [31, 32]. In summary, we find supportive evidence for the occurrence of substrain variants by endogenous reactive oxygen species (ROS); a common source of oxidative DNA damage [33]. Based on the mutational spectrum, non-negative matrix factorization (NMF) can be used to identify more detailed underlying mutational signatures. However, when we carry out such analyses we do not find a significant difference in mutational signature between substrain and control SNVs, suggesting that substrain variation results from common mutational processes and thus represents ongoing evolutionary processes.

Next, we investigated the functional consequences of the 28,318 substrain variants by analyzing the nonsynonymous to synonymous ratio, which we previously used as a measure of selective pressure. Interestingly, we find relatively more nonsynonymous SNVs in the substrain variants compared to the control set, indicating that the substrain SNVs more often affect protein sequence (Figure 5b). To substantiate this finding and to get a gene annotation-independent measure of the functional impact of the substrain variants, we also retrieved the phastCons scores [34] per variant. This score (between 0 and 1) is calculated for each nucleotide in the genome as a measure for evolutionarily constraint and was derived by comparing the rat genome to 8 other species: mouse, dog, cow, opossum, chicken, frog, zebrafish and human. In line with the previous results we find a significantly higher phastCons score of the substrain-affected nucleotides compared to the control set ($p<0.0001$; Figure 5c). These two lines of evidence suggest that evolutionary pressure has not (yet) selected against these possibly damaging variants, confirming the relatively young age of the substrain variants. On the other hand, substrain-specific variants may

have a relatively large effect on protein function and thus on associated biology and it is therefore extremely important to know this category of variation when comparing experimental results obtained with different rat substrains in different labs.

Discussion

Although RGSC5.0 was already released in 2012, all whole-genome sequencing studies to date are based on the much older RGSC3.4 assembly. Here, we merge publicly available whole genome sequence data of 40 widely used rat inbred strains and substrains into a comprehensive integrated variant inventory. This resource allows researchers to functionally annotate their data on the more recent RGSC5.0.

Integrated analysis of a large number of strains increases effective genomic coverage and thus improves on variant calling sensitivity. The multi-sample variant calling approach used here makes optimally use of this [7-12], resulting in a more accurate and more complete set of called variants, especially in strains with lower coverage at a given position. The resulting resource is useful for a broad range of researchers who use rats for studying genetic traits and can easily be exploited. For example, this inventory can be used for choosing strains and substrains for specific experiment or as controls, when knowing their genetic differences in a locus of interest. Another way to use this resource is by coupling it to Quantitative Trait Locus (QTL) data, which is available for many of these strains for a broad range of complex traits [8-10, 12, 35-39]. This allows for filtering for shared and unique variants between strains with and without the trait to narrow in on potential causal variants. Finally, the resource can be used for strain of origin designation when WGS or genotyping data is available, as exemplified by the SUO_F344 WGS data included in this study.

We showed that the biological relevance of most high impact SNVs is limited. In part, this effect can be attributed to the low expression level of the gene or to skipping of exons in which a high impact variant is found. Furthermore, a small part of the automatically predicted deleterious variants appeared false positives caused by the lack of taking neighboring variants into account in the effect prediction. Addressing this effect

requires adaptions of the current effect prediction calling algorithms. When we investigate the population structure of the 40 rat strains, we find a distinction between nine separated clusters, which recapitulates the previously published origin of some of these strains [11]. We see that the genomic variant distribution in more than 65% of the strains (27 out of 40) has a clearly distinct pattern between clusters. In addition, all strains in cluster 9 show a confetti-like genomic distribution of multiple clusters, possibly reflecting their heterogeneous, yet shared, origin. Similar to data from mice [40] we observe introgression of shared haplotypes between strains, suggesting intercrosses in rat strain selection processes. Using SNP marker information in rat, it was already shown that this effect was present [5, 6] and here we confirm this observation on a genome-wide scale. Furthermore, we identified substrain variation in seven rat strains and find that the degree of variation is highly variable between strains. The strain with the highest degree of substrain variation is WKY and part of this variation can be explained by their distribution to different geographical locations before complete inbreeding [25]. When we further investigate the different aspects of substrain variants we can explain part of their origin by introgression and part by ongoing evolution through genetic drift. In general the characteristics of substrain variants matches with their recent origin. Firstly the impact of the substrain variants is relatively high. Secondly the substrain variants show evidence for endogenous ROS DNA damage, a process that continuously challenges the integrity of DNA [33].

In summary, we present a comprehensive inventory of uniformly called genomic variants mapped on the RGSC5.0 reference assembly for a range of commonly used inbred rat strains. This resource is valuable for a broad range of researchers that use rats in biomedical and complex genetics research and may facilitate further research on rat functional genomics and interspecies comparison. The knowledge on substrain variation may assist experimental design and improve on the outcome and reproducibility of experimental results between institutes and thus improve the overall quality of biomedical animal research.

Methods

Genome and transcriptome sequencing

We performed whole genome sequencing on the rat strains: BN-*Lx*/CubPrin, SHR/OlaIpcvPrin, SHR/NCrlPrin, and SUO_F344. All animals were obtained from stock maintained by Dr. Morton Printz, Department of Pharmacology, University of California San Diego. Genomic DNA was extracted from 25 mg of homogenized cortical tissue using the DNeasy Blood and Tissue kit (#69504, Qiagen). One microgram of genomic DNA was used as input in the Illumina TruSeq DNA Kit (#PE-940-2001, Illumina) following the manufacturer's instructions. The libraries were sequenced using 100 cycles paired-end reads on an Illumina HiSeq2000 following the manufacturer's instructions. We performed RNA sequencing on a male BN-*Lx*/Cub of snap-frozen and powdered whole tissues. Total RNA from heart, muscle and skin was isolated was firstly isolated using the TRIzol® reagent (#15596-026, Invitrogen, Life Technologies). After this total RNA was (re)isolated using the Promega Maxwell® 16 MDx Research System (#AS3000, Promega) with the Maxwell® 16 LEV simplyRNA Blood Kit (#AS1310, Promega) for brain, heart, kidney, liver, lung, muscle, ovary, skin, spleen, testis, thymus and whole blood. One microgram of isolated total RNA was used as input for sample prep using TruSeq Stranded Total RNA Kit with Ribo Zero Human/Mouse/Rat (#RS-122-2203, Illumina) following the manufacturer's instructions. The libraries were sequenced 101 cycles paired-end in rapid run modus on an Illumina HiSeq2500 following standard manufacturer's instructions.

Mapping, variant calling and annotation

For the whole genome sequencing data the 32 strains that were sequenced on Illumina platforms were mapped with BWA mem –M 0.7.5a [43]. The 10 strains that were sequenced on SOLiD platform were mapped with BWA 0.5.9 aln -c -l 25 -k 2 -n 10 (the latest version to support color space). Picard MarkDuplicates version 1.89 was used to mark all the duplicate reads per rat strain. SNV and indel calling was done following the GATK HaplotypeCaller v2.8-1-g932cd3a best practices from the Broad Institute [16]. SNVs and indels were annotated using SnpEff version 3.3h [17]. Structural variant calling was done using DELLY version 0.3.3 with -q 20 [26] and CNVnator version 0.2.7 with a bin size of 1,000 bp [27].

RNA sequencing downstream analysis

For the RNA sequencing of BN-*Lx*/Cub tissues, reads were mapped to the genome first to detect and remove sequences with multiple alignments. The remaining sequences were then aligned with TopHat 1.4.1 [44] against the RGSC5.0 reference genome and transcriptome based on Ensembl gene annotations [45]. To align reads across both novel and known splice junctions, we also allowed the discovery of unknown splice junctions. We then counted uniquely aligning reads that could be assigned unambiguously to one gene. This count data was then normalized for gene length and library size to obtain genome-wide FPKM values. 'Percent Spliced In' (PSI) values were generated by counting reads either mapping into (inclusion read) or jumping over (exclusion read) a given exon. After length normalization, the ratio between inclusion reads was divided by the sum of inclusion and exclusion reads to obtain the PSI score for each exon. As a control, a set of 16,000 randomly chosen exons was taken. A PSI value of 1 indicates constitutive exons, whereas values below 1 show exons that are not present in every transcript. Only exons of expressed genes (FPKM >= 1) were considered. If neither inclusion nor exclusion reads were present, a PSI value of 0 was assigned to indicate that the exon was not used.

Downstream genomic variant analysis

Cross-species comparison

Next to the rat data described in this paper, we used variomes of dog (assembly canFam3), horse (equCab2), mouse (NCBIM37/Mm9) and pig (susScr3/Sscrofa10.2). Corresponding variants and genome sequences were downloaded from Ensembl database (release 75, ftp://ensembldb.ensembl.org). Variants from each of these species were transposed to mouse genome NCBIM38/Mm10 using corresponding UCSC Chain alignments. Number of polymorphic positions was calculated for sliding windows (containing 100kb syntenic sequence, 25 Kb step between starting position of adjacent windows). Z-score transformed values were used for plotting the regions where: 1) all species showed low level of variation, i.e. were all in lower 10 percentiles. 2) all species showed high level of variation, i.e. were all in upper 10 percentiles.

fastStructure

We used all homozygous variants as input in the fastStructure algorithm [24] (http://pritchardlab.stanford.edu/structure.html). We determined the population structure for K=2 until K=31 and determined the appropriate number of model components that explain structure in the dataset by running the build-in script chooseK.py. In order to determine the genomic distributions of these clusters we divided the genome in segments containing 20.000 SNVs, in each window the genotypes of the different rat strains were compared to the average genotype profile of each of the 9 groups. Similarity scores were calculated using Spearman correlation; each window was assigned a group membership based on the maximum correlation coefficient.

phastCons

Conservation scores for alignments of 8 vertebrate genomes with Rat (PhastCons9way scores [34], rn4 assembly (Nov. 2004)) were downloaded from UCSC Genome browser FTP server. Since no phastCons scores were available yet for the RGSC5.0 assembly, UCSC LiftOver was used to retrieve the new coordinates of phastCons scores.

Supplemental information

All supplemental files in this chapter can be downloaded from http://www.hubrecht.eu/research/cuppen/hermsen_thesis.html

Availability of supporting data

The genome sequence data for the four rat strains (BN-*Lx*/CubPrin, SHR/OlaIpcvPrin, SHR/NCrlPrin, and SUO_F344), supporting the results of this article, is available in the European Bioinformatics Institute (EBI) Short Read Archive (SRA) under accession [EBI-SRA: PRJEB6956]. The BN-Lx/Cub RNA-seq data, supporting the results of this article, is available in the European Bioinformatics Institute (EBI) Short Read Archive (SRA) under accession [EBI-SRA: PRJEB6938]. SNVs, indels and structural variants in all 40 strains are available by browsing via the Rat Genome Database (http://rgd.mcw.edu/) or via a direct download of the VCF file per variant type: ftp://ftp.rgd.mcw.edu/pub/strain_specific_variants/Hermsen_et_al_40Genomes_Variants/. In addition data from the four newly sequenced strains is also available via PhenoGen Informatics (http://phenogen.ucdenver.edu).

Authors' contributions

RH and EC conceptually designed the study, critically discussed results and wrote the manuscript. TA and NH provided sequencing data for most strains in this study. SF and BT generated the newly presented next-generation sequencing data for four rat strains. WS and SB performed next-generation sequencing mapping and variant calling. JL, FB, SS, EA, RB, RHW, MS, VG and EC contributed to scientific discussions and data analysis. All authors read and approved the final version of the manuscript.

Acknowledgements

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). WS was financially supported by the Netherlands Bioinformatics Centre (NBIC). The work performed by SF and BT was supported by grants from NIH/NIAAA 5 T32AA007464-38 and 5 R24AA013162-13. We are grateful to Dr. Morton P. Printz (University of California, Department of Pharmacology, San Diego) for kindly providing tissue from four rat strains for DNA sequencing.

References

1. Jacob HJ: Functional genomics and rat models. Genome research 1999, 9(11):1013-1016.
2. Bosse JD, Lin HY, Sloan C, Zhang QJ, Abel ED, Pereira TJ, Dolinsky VW, Symons JD, Jalili T: A low-carbohydrate/high-fat diet reduces blood pressure in spontaneously hypertensive rats without deleterious changes in insulin resistance. American journal of physiology Heart and circulatory physiology 2013, 304(12):H1733-1742.
3. Diness JG, Skibsbye L, Jespersen T, Bartels ED, Sorensen US, Hansen RS, Grunnet M: Effects on atrial fibrillation in aged hypertensive rats by Ca(2+)-activated K(+) channel inhibition. Hypertension 2011, 57(6):1129-1135.
4. Sagvolden T, Dasbanerjee T, Zhang-James Y, Middleton F, Faraone S: Behavioral and genetic evidence for a novel animal model of Attention-Deficit/Hyperactivity Disorder Predominantly Inattentive Subtype. Behavioral and brain functions : BBF 2008, 4:56.
5. Saar K, Beck A, Bihoreau MT, Birney E, Brocklebank D, Chen Y, Cuppen E, Demonchy S, Dopazo J, Flicek P et al: SNP and haplotype mapping for genetic analysis in the rat. Nature genetics 2008, 40(5):560-566.
6. Smits BM, Guryev V, Zeegers D, Wedekind D, Hedrich HJ, Cuppen E: Efficient single nucleotide polymorphism discovery in laboratory rat strains using wild rat-derived SNP candidates. BMC genomics 2005, 6:170.
7. 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 research 2010, 20(6):791-803.
8. 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 biology 2012, 13(4):r31.

9. 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. Nature genetics 2013, 45(7):767-775.

10. Guo X, Brenner M, Zhang X, Laragione T, Tai S, Li Y, Bu J, Yin Y, Shah AA, Kwan K et al: Whole-genome sequences of DA and F344 rats with different susceptibilities to arthritis, autoimmunity, inflammation and cancer. Genetics 2013, 194(4):1017-1028.

11. Atanur SS, Diaz AG, Maratou K, Sarkis A, Rotival M, Game L, Tschannen MR, Kaisaki PJ, Otto GW, Ma MC et al: Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. Cell 2013, 154(3):691-703.

12. Ma MC, Atanur SS, Aitman TJ, Kwitek AE: Genomic structure of nucleotide diversity among Lyon rat models of metabolic syndrome. BMC genomics 2014, 15:197.

13. 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(6982):493-521.

14. van Heesch S, Kloosterman WP, Lansu N, Ruzius FP, Levandowsky E, Lee CC, Zhou S, Goldstein S, Schwartz DC, Harkins TT et al: Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing. BMC genomics 2013, 14:257.

15. The Rat Genome Project.https://http://www.hgsc.bcm.edu/other-mammals/rat-genome-project.

16. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 2010, 20(9):1297-1303.

17. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 2012, 6(2):80-92.

18. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C et al: Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature genetics 2011, 43(10):956-963.

19. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 2009, 4(1):44-57.

20. Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar A, Lindblad-Toh K: The genomic signature of dog domestication reveals adaptation to a starch-rich diet. Nature 2013, 495(7441):360-364.

21. Niimura Y: Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. Human genomics 2009, 4(2):107-118.

22. Hardison RC: Evolution of hemoglobin and its genes. Cold Spring Harbor perspectives in medicine 2012, 2(12):a011627.

23. Higashino A, Sakate R, Kameoka Y, Takahashi I, Hirata M, Tanuma R, Masui T, Yasutomi Y, Osada N: Whole-genome sequencing and analysis of the Malaysian cynomolgus macaque (Macaca fascicularis) genome. Genome biology 2012, 13(7):R58.

24. Raj A, Stephens M, Pritchard JK: fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics 2014, 197(2):573-589.

25. Kurtz TW, Montano M, Chan L, Kabra P: Molecular evidence of genetic heterogeneity in Wistar-Kyoto rats: implications for research with the spontaneously hypertensive rat. Hypertension 1989, 13(2):188-192.

26. Rausch T, Zichner T, Schlattl A, Stutz AM, Benes V, Korbel JO: DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 2012, 28(18):i333-i339.

27. Abyzov A, Urban AE, Snyder M, Gerstein M: CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome research 2011, 21(6):974-984.

28. Kren V: Genetics of the polydactyly-luxate syndrome in the Norway rat, Rattus norvegicus. Acta Universitatis Carolinae Medica Monographia 1975(68):1-103.

29. Wotjak CT: C57BLack/BOX? The importance of exact mouse strain nomenclature. Trends in genetics : TIG 2003, 19(4):183-184.

30. Denver DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledo JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M et al: A genome-wide view of Caenorhabditis elegans base-substitution mutation processes. Proceedings of the National Academy of Sciences of the United States of America 2009, 106(38):16310-16314.

31. Tomita-Mitchell A, Kat AG, Marcelino LA, Li-Sucholeiki XC, Goodluck-Griffith J, Thilly WG: Mismatch repair deficient human cells: spontaneous and MNNG-induced mutational spectra in the HPRT gene. Mutation research 2000, 450(1-2):125-138.

32. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL et al: Signatures of mutational processes in human cancer. Nature 2013, 500(7463):415-421.

33. Evans MD GH, Lunec J: Reactive oxygen species and their cytotoxic mechanisms. Advances in Molecular and Cell Biology 1997, 20:25–73.

34. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research 2005, 15(8):1034-1050.

35. Lopez B, Ryan RP, Moreno C, Sarkis A, Lazar J, Provoost AP, Jacob HJ, Roman RJ: Identification of a QTL on chromosome 1 for impaired autoregulation of RBF in fawn-hooded hypertensive rats. American journal of physiology Renal physiology 2006, 290(5):F1213-1221.

36. Rapp JP, Garrett MR, Dene H, Meng H, Hoebee B, Lathrop GM: Linkage analysis and construction of a congenic strain for a blood pressure QTL on rat chromosome 9. Genomics 1998, 51(2):191-196.

37. Vanderlinden LA, Saba LM, Printz MP, Flodman P, Koob G, Richardson HN, Hoffman PL, Tabakoff B: Is the Alcohol Deprivation Effect Genetically Mediated? Studies with HXB/BXH Recombinant Inbred Rat Strains. Alcoholism, clinical and experimental research 2014, 38(7):2148-2157.

38. Yagil Y, Hessner M, Schulz H, Gosele C, Lebedev L, Barkalifa R, Sapojnikov M, Hubner N, Yagil C: Geno-transcriptomic dissection of proteinuria in the uninephrectomized rat uncovers a molecular complexity with sexual dimorphism. Physiological genomics 2010, 42A(4):301-316.

39. Zagato L, Modica R, Florio M, Torielli L, Bihoreau MT, Bianchi G, Tripodi G: Genetic mapping of blood pressure quantitative trait loci in Milan hypertensive rats. Hypertension 2000, 36(5):734-739.

40. Yang H, Wang JR, Didion JP, Buus RJ, Bell TA, Welsh CE, Bonhomme F, Yu AH, Nachman MW, Pialek J et al: Subspecific origin and haplotype diversity in the laboratory mouse. Nature genetics 2011, 43(7):648-655.

41. Laulederkind SJ, Hayman GT, Wang SJ, Smith JR, Lowry TF, Nigam R, Petri V, de Pons J, Dwinell MR, Shimoyama M et al: The Rat Genome Database 2013--data, tools and users. Briefings in bioinformatics 2013, 14(4):520-526.

42. Bennett B, Saba LM, Hornbaker CK, Kechris KJ, Hoffman P, Tabakoff B: Genetical genomic analysis of complex phenotypes using the PhenoGen website. Behavior genetics 2011, 41(4):625-628.

43. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25(14):1754-1760.

44. Trapnell C, Pachter L, Salzberg SL: TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009, 25(9):1105-1111.

45. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al: Ensembl 2014. Nucleic acids research 2014, 42(Database issue):D749-755.

3

Chapter 3

Combined sequence-based and genetic mapping analysis of complex traits in outbred rats

Rat Genome Sequencing and Mapping Consortium

Amelie Baud[1], Roel Hermsen[2], Victor Guryev[3,2], Pernilla Stridh[4], Delyth Graham[5], Martin W. McBride[5], Tatiana Foroud[6], Sophie Calderari[7], Margarita Diez[4], Johan Ockinger[4], Amennai D. Beyeen[4], Alan Gillett[4], Nada Abdelmagid[4], Andre Ortlieb Guerreiro-Cacais[4], Maja Jagodic[4], Jonatan Tuncel[8], Ulrika Norin[8], Elisabeth Beattie[5], Ngan Huynh[5], William H. Miller[5], Daniel L. Koller[6], Imranul Alam[9], Samreen Falak[10], Mary Osborne-Pellegrin[11], Esther Martinez-Membrives[12], Toni Canete[12], Gloria Blazquez[12], Elia Vicens-Costa[12], Carme Mont-Cardona[12], Sira Diaz-Moran[12], Adolf Tobena[12], Oliver Hummel[10], Diana Zelenika[13], Kathrin Saar[10], Giannino Patone[10], Anja Bauerfeind[10], Marie-Therese Bihoreau[13], Matthias Heinig[14,10], Young-Ae Lee[10,15], Carola Rintisch[10], Herbert Schulz[10], David A. Wheeler[16], Kim C. Worley[16], Donna M. Muzny[16], Richard A. Gibbs[16], Mark Lathrop[13], Nico Lansu[2], Pim Toonen[2], Frans Paul Ruzius[2], Ewart de Bruijn[2], Heidi Hauser[17], David J. Adams[17], Thomas Keane[17], Santosh S. Atanur[18], Tim J. Aitman[18], Paul Flicek[19], Tomas Malinauskas[20], E. Yvonne Jones[20], Diana Ekman[8], Regina Lopez-Aumatell[12], Anna F Dominiczak[5], Martina Johannesson[8], Rikard Holmdahl[8], Tomas Olsson[4], Dominique Gauguier[7], Norbert Hübner[21,10], Alberto Fernandez-Teruel[12], Edwin Cuppen[2], Richard Mott[1], Jonathan Flint[1]

adapted from Nature Genetics 2013 45:767-775

Author affiliations

1 Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford, OX3 7BN, UK

2 Hubrecht Institute, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands

3 European Research Institute for the Biology of Ageing, RuG, UMCG, Antonius Deusinglaan 1, 9713 AV, Groningen, The Netherlands

4 Neuroimmunology Unit. Department of Clinical Neuroscience, Karolinska Institutet, CMM L8:04, 17176 Stockholm

5 BHF Glasgow Cardiovascular Research Centre, Institute of Cardiovascular & Medical Sciences, Glasgow University, 126 University Place, Glasgow, G12 8TA, UK

6 Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana USA

7 INSERM UMRS872, Cordeliers Research Centre, 15 rue de l'Ecole de Medecine, 75006 Paris, France

8 Division of Medical Inflammation Research, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, 171 77 Stockholm, Sweden

9 Department of Orthopaedic Surgery, Indiana University School of Medicine, Indianapolis, Indiana USA

10 Max-Delbruck Center for Molecular Medicine, Berlin D-13092, Germany

11 Inserm U698, Hôpital Bichat, Paris, France

12 Medical Psychology Unit, Department of Psychiatry & Forensic Medicine, Institute of Neurosciences, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

13 Commissariat à l'énergie Atomique, Institut de Génomique, Centre National de Génotypage, Evry, France

14 Department of Computational Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

15 Pediatric Allergology, Experimental and Clinical Research Center, Charité Universitätsmedizin Berlin, Germany

16 Human Genome Sequencing Center, One Baylor Plaza, MSC-226, Houston, TX 77030

17 The Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK.

18 Physiological Genomics and Medicine Group, Medical Research Council Clinical Sciences Centre, Faculty of Medicine, Imperial College London, Hammersmith Hospital, London W12 0NN, United Kingdom

19 European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, United Kingdom

20 Division of Structural Biology, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

21 DZHK (German Centre for Cardiovascular Research), partner site Berlin, Berlin, Germany

Abstract

Genetic mapping on fully sequenced individuals is transforming our understanding of the relationship between molecular variation and variation in complex traits. Here we report a combined sequence and genetic mapping analysis in outbred rats that maps 355 quantitative trait loci for 122 phenotypes. We identify 31 causal genes involved in 27 phenotypes, implicating novel genes in models of anxiety, heart disease and multiple sclerosis. The relation between sequence and genetic variation is unexpectedly complex: at approximately 40% of quantitative trait loci a single sequence variant cannot account for the phenotypic effect. Using comparable sequence and mapping data from mice, we show the extent and spatial pattern of variation in inbred rats differ significantly from those of inbred mice, and that the genetic variants in orthologous genes rarely contribute to the same phenotype in both species.

Introduction

Unraveling the complex relationship between phenotype and genotype poses a formidable challenge for biomedical science. Despite considerable success in identifying genetic loci that contribute to quantitative variation and disease susceptibility in humans [1], in most organisms the causal genetic variants at loci that contribute to complex phenotypes remain unclear [2]. Finding the responsible molecular changes would allow us to understand how phenotypic variation arises and confirm the identity of relevant genes. In this report we present results from an outbred rat heterogeneous stock (hereafter NIH-HS) in a combined sequence-based and genetic mapping analysis of 160 phenotypes. The NIH-HS, established in the 1980s in NIH, is descended from eight inbred progenitors, BN/SsN, MR/N, BUF/N, M520/N, WN/N, ACI/N, WKY/N, and F344/N [3], containing segregating variation representative of commonly used laboratory rats. Heterogeneous stocks (HS) have three characteristics suited to genetic mapping: (i) quantitative trait loci (QTLs) can be resolved to megabase resolution; (ii) the complete sequence of genotyped HS animals can be imputed with high accuracy from the progenitor genomes; (iii) the population has a well-defined haplotype space that can be exploited to determine whether genetic association is caused by single sequence

variants or by haplotypes [4-6]. This distinction is fundamental to understanding the signals from genome wide association studies, where it is unknown how often causality can be attributed to a single variant. In natural populations it is rarely feasible to test for haplotypic effects because of the difficulty of estimating the large number of unknown rare haplotypes [7]. Here we describe the sequence of the eight progenitors, the development of a rat SNP array, the genotyping and phenotyping of 1,407 outbred NIH-HS rats, and the mapping of hundreds of quantitative trait loci (QTLs). We use the haplotypic properties of the NIH-HS to investigate the molecular basis of these QTLs.

Results

Sequence analysis

We generated sequence data equivalent to an average of 22X SOLiD coverage of the eight NIH-HS inbred founder strains. After mapping to the reference strain (BN/NHsdMcwi [8]) we report our results with respect to an accessible genome, which represents ~88% of the reference genome (Table 1). We identified 7.2M SNP sites (containing 19.8M genotypes differing from the reference in at least one strain), 633,000 indels (<10bp with the majority consisting of one (79.3%) or two (12.3%) basepair changes) and 44,000 structural variants. We assessed the sensitivity and specificity of variant calls by comparison with 2.1 Mb of DNA from one non-reference strain, LE/Stm, finished to an estimated accuracy of one error per 100,000 bp [9]. Although LE/Stm is not an NIH-HS progenitor strain, it is one of the few non-reference rat strains cloned into a library of bacterial artificial chromosomes (BACs) (so suitable for highly accurate clone based sequencing)[9] and one that is similarly diverged from the reference strain (BN/NHsdMcwi). Comparison of SOLiD and capillary variant calls showed 2.7% of SNPs, 2.2% of indels and 16.7% of structural variants were false positive calls. These error rates were independently confirmed in the NIH-HS strains by analysis of a randomly selected subset of variants using PCR-based resequencing, which confirmed all selected SNPs (84/84) and indels (80/80) and most of the called structural variants (53/54). In contrast, false-negative rates were much higher: 17.2% for SNPs, 41.4% for indels and 65% for structural variants. Most false-negative SNPs and indels are next to repeats (77.9 and 80.8% respectively). Table 1 summarises the

variation in each strain. Excluding BN/SsN (which is a sub-strain of the reference, with consequently far fewer differences than the other strains), the average number of SNPs per strain is 2.8M.

Strain	Gb of mapped data	Coverage	% of genome inaccessible	SNPs	Private SNPs	Indels	Private indels	Structural variants	Private structural variants
ACI/N	65.9	26.3	12.6	2,883,405	228,468	166,425	12,646	19,499	756
BN/SsN	54.4	21.7	9.4	71,038	563,308	0	14,839	27	4,203
BUF/N	62.3	24.9	12.7	2,748,633	125,202	172,934	7,195	22,176	1,002
F344/N	77.9	31.1	11.8	2,831,144	97,951	157,522	5,007	25,257	1,003
M520/N	72.5	28.9	12.3	2,836,898	89,277	170,031	5,008	24,090	915
MR/N	62.4	24.9	12.3	2,664,124	223,514	151,099	12,005	18,306	1,004
WKY/N	63.4	25.3	12.1	3,088,953	496,327	164,634	23,979	28,270	3,357
WN/N	62.3	24.9	12.2	2,698,493	249,563	154,769	13,541	18,563	700

Table 1 Sequence variation in the eight progenitor strains of NIH-HS rats

Nucleotide diversity in NIH-HS progenitors

Sequence diversity in the NIH-HS progenitors has the following characteristics. First, diversity between all pairs of strains is similar, so that there are no extremely sequence divergent strains (Supplementary Figure 1). Second, in total 29% of 7.2M SNPs are private to a strain, hence unique haplotypes are relatively common in the NIH-HS. Third, regions of low diversity are small (median 400kb), with no blocks over 35 Mb (Figure 1a). Within divergent regions, there is a median of 151 differences per 100kb (Figure 1b). In comparison with the eight inbred strains that founded the mouse HS [4, 10], the rat founders are less diverse (10.2M SNPs in the mouse founders), but that diversity is more homogeneous: in the mouse genomes, long tracts of identical haplotypes alternate with segments of much greater diversity (Figures 1a, 1b).

Phenotype	Disease model	Number of measures	Week
Coat colour		4	7
Wound healing		1	7 and 17
Fear related behaviours	Anxiety	10	8 to 10
Glucose tolerance	Type II diabetes	6	11
Cardiovascular function	Hypertension	2	12
Body weight	Obesity	1	13
Basal hematology		26	13
Basal immunology		34	13
Induced neuroinflammation	Multiple sclerosis	11	13 to 17
Bone mass and strength	Osteoporosis	43	17
Arterial elastic lamina ruptures		6	17
Serum biochemistry		15	17
Renal agenesis		1	17

Table 2: Summary of phenotypes collected.

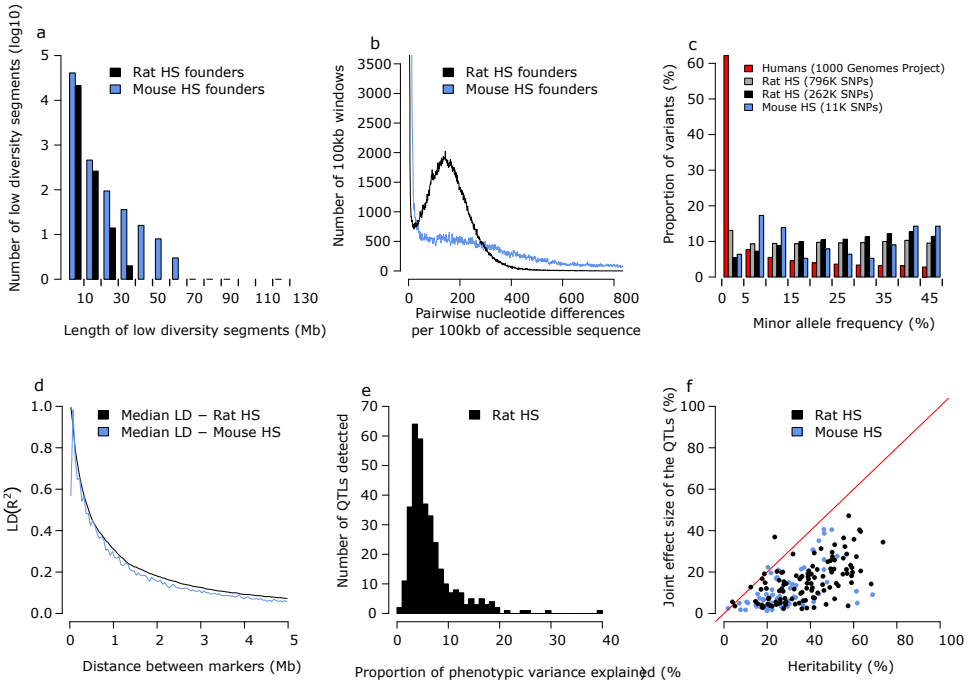


Figure 1 Sequence diversity among progenitor strains and genetic architecture of the rat NIH-HS. a) Regions of low diversity in the rat (black) and mouse (blue) progenitors. The horizontal axis shows the length in megabases of genomic regions with little sequence divergence (less than 13 SNPs/100kb). The vertical axis shows the numbers of segments observed in the eight progenitors. b) Sequence divergence in the progenitors. The horizontal axis is a measure of pairwise sequence diversity, the number of sequence differences observed in windows of 100 kilobases, the vertical axis gives the number of observations. The horizontal axis is truncated at 800 sequence differences and the vertical axis at 3500 windows. c) Minor allele frequencies in rat (gray & black), mouse (blue) and human (red) populations. The rat analysis was performed with the set of autosomal

markers used to reconstruct haplotypes (261,684) as well as the complete set of 796,187 autosomal variants on the RATDIV array. d) The extent of linkage disequilibrium (measured as R2) in the rat NIH-HS. Distances between pairs of autosomal markers were binned (horizontal axis). The vertical axis shows the median of the corresponding distribution of LD. e) The distribution of effect sizes for the 343 loci mapped by mixed models in the rat NIH-HS. The horizontal axis is the proportion of phenotypic variance attributable to each locus. f) The proportion of heritability that can be explained by the joint effect of the QTLs detected for each phenotype. Each dot represents a single phenotype, with the horizontal axis showing the heritability and the vertical axis the joint QTL effect for that phenotype.

Phenotypes and genotypes

The NIH-HS rats were phenotyped with a protocol that includes six disease models (anxiety, diabetes, hypertension, aortic elastic lamina rupture, multiple sclerosis, osteoporosis) and measures of risk factors for common diseases (e.g. lipid and cholesterol levels and cardiac hypertrophy) [11] (Table 2). In total, 160 phenotypes were measured (Supplementary Table 1). We selected 1,407 animals for genotyping and 198 non-phenotyped parents, together with the HS founders. We designed a high density Affymetrix SNP genotyping array (RATDIV), using sequences from 13 inbred strains, which interrogated 803,485 SNPs. The SOLiD and RATDIV calls agreed at 99.98% of the 560,000 SNPs segregating in the 8 NIH-HS founders. We genotyped the NIH-HS with the array, and reconstructed the mosaics of NIH-HS founder haplotypes from 265,551 polymorphic high quality SNPs. In the NIH-HS the mean minor allele frequency is 22% (Figure 1c) and linkage disequilibrium falls below 0.2 (median r^2) within 1Mb on the autosomes (Figure 1d). Four pairs of loci show high interchromosomal LD, due to mis-assembly of the reference sequence used here (Rnor3.4); these loci were excluded from the analysis (Supplementary Table 2).

Quantitative Trait Loci

The NIH-HS contains individuals of varying relatedness that generate population structure and hence false positive genetic associations. We evaluated two strategies for dealing with relatedness; mixed models in which the genotypic similarity matrix between individuals modeled their phenotypic correlation [12], and resampling methods to identify loci that replicate consistently across multiple QTL models fitted on subsamples of the mapping population [13]. In both strategies, QTLs were detected by haplotype association [14]. We compared the methods by simulation to find out which best controlled the false positive rate while retaining power. Mixed models performed better than resampling when phenotypes

were simulated to be normally distributed, but the reverse was true for non-normally distributed phenotypes (i.e. binary phenotypes and those with a negative binomial distribution). Since these methods have different advantages, we mapped all traits with both methods, but only report those QTLs detected at false discovery rate (FDR) of 10% by that method which performed best for each trait (thresholds in Supplementary Table 1). Figure 2 shows a genome scan for one phenotype (platelet aggregation), revealing three loci at 10% FDR.

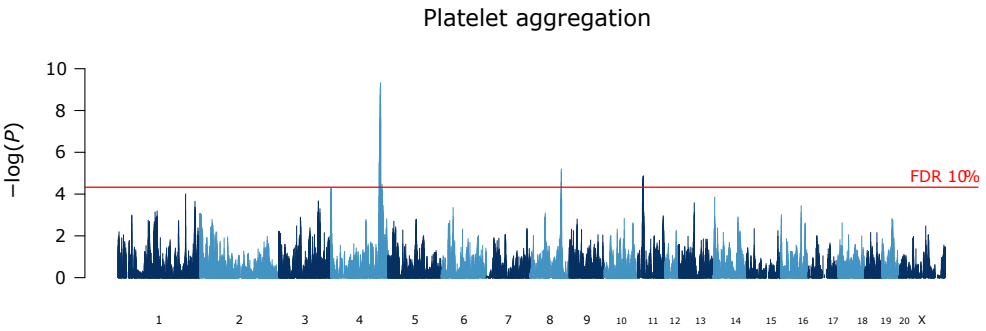


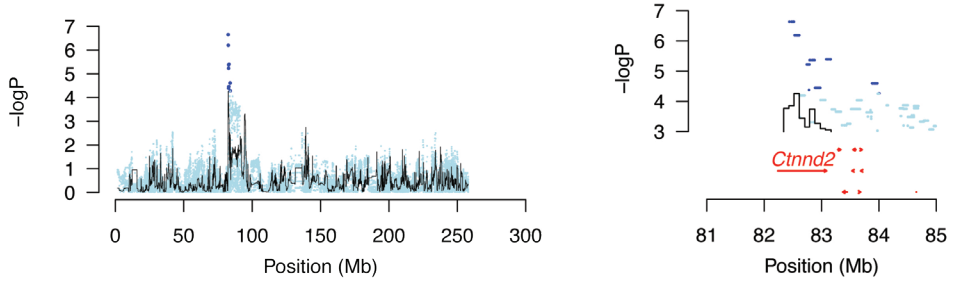
Figure 2 Genome scan for platelet aggregation. The scan shows the results of a haplotype mixed model. The *y* axis scale shows the negative log *P* values for association with variation in platelet aggregation. The association peak on chromosome 4 harbors the von Willebrand factor gene that is identified through sequence analysis as the causative gene.

We identified 355 QTLs for 122 phenotypes with a mean of 2.9 QTLs per phenotype (Supplementary Table 3). The number of QTLs per phenotype and the QTL effect sizes (Figure 1e) have markedly skewed distributions, with a median effect size of 5% (mean effect size 6.5%). Large effect QTLs are rare: only 22 QTLs explain more than 15% of the variance. We identified 28 QTLs that explained less than 2.5% of the phenotypic variance. Figure 1f shows the correlation between the heritability and the total variance explained jointly by the detected QTLs. On average the QTLs explain 42% of the heritable phenotypic variance. In comparison with QTLs mapped in other rat crosses in the Rat Genome Database, there is significant overlap with NIH-HS QTLs for the number of arterial elastic lamina ruptures, total cholesterol levels and heart weight (at a nominal *P* value of 0.05; Supplementary Table 4). We estimated the QTLs' confidence intervals by simulating a large number of QTLs throughout the genome with various effect sizes, and calculated the distribution of the confidence intervals' widths as a function of their significance (Supplementary Figure 2). The median size of the 90% confidence interval is 4.5Mb, on average containing more than 40 genes.

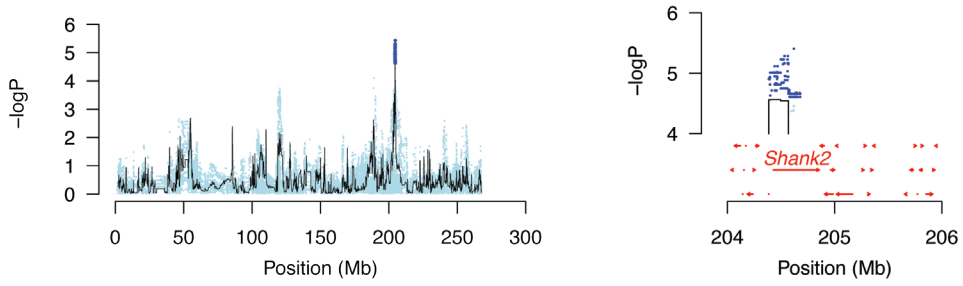
Incorporation of sequence with mapping data

We investigated the extent to which our near complete catalogue of segregating sequence variants would identify genes and causative mutations. The HS permits a test, called merge analysis [6], of whether a variant is responsible for phenotypic variation, under the assumption that a single imputed variant, or variants on a single progenitor haplotype, are causal. Because genetic variation segregates in the form of the progenitor haplotypes in the HS, QTLs can always be explained by variation in the haplotypes. When a QTL is due to a single variant though, genotypic variation at the variant will explain phenotypic variation better than progenitor haplotypes. To measure whether a single variant explains the QTL we calculated the difference d = log P_{merge} – log $P_{haplotype}$ where log $P_{haplotype}$, is the maximum negative $\log_{10} P$ value of the haplotype test of no association and log P_{merge} is the maximum of all merge $\log_{10} P$ values of variants under the QTL. Any imputed variant with a merge $\log_{10} P$ value that exceeds the maximum haplotype $\log_{10} P$ value was termed a candidate variant. If d was <0 then no candidate variants exist at the QTL. We investigated the characteristics of these candidate variants at 343 QTLs mapped using mixed models: at 131 QTLs (38%) we identified at least one candidate variant (Supplementary Table 3). There are three ways in which focusing on these candidate variants helps identify genes at a QTL. First, we increase resolution by ruling out the great majority (usually over 90%) of sequence variants under most QTLs as being causal. We found 28 QTLs at which only a single gene contained candidate variants (Table 3). An example is Catenin-delta 2 (*Ctnnd2*) at a QTL for an anxiety-related phenotype (Figure 3a). CTNND2 is a protein found in complexes with cadherin cell adhesion molecules at neuronal synapses [15]. Figure 3b shows another example for a locus influencing heart weight, where out of 82 coding genes under the QTL, only *Shank2* contained candidate SNPs. *Shank2* encodes a synaptic protein [16] not previously associated with cardiovascular physiology. Second, merge analysis identifies some candidate variants lying within coding regions. Those predicted to affect protein structure are more likely to be causal. Thus we identified a potential causal nucleotide in a QTL for antibody recognition of CD45RC on CD4+ and CD8+ T cells (Figure 3c). The antibody used binds to the CD45RC isoform, which expresses a C-domain, encoded by the sixth exon, in which we found a candidate variant changing an amino acid (p.Arg114His).

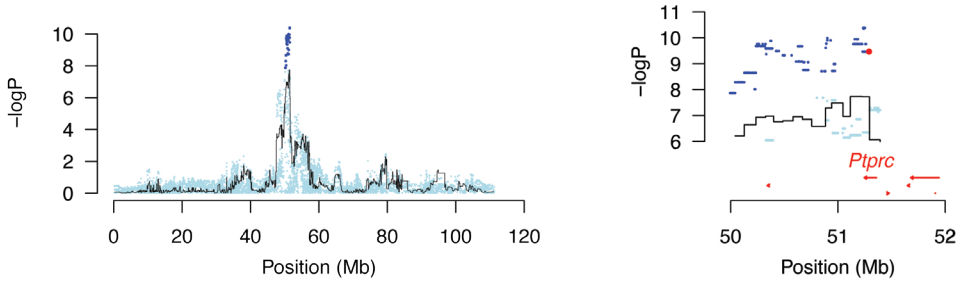
a Anxiety (chromosome 2)



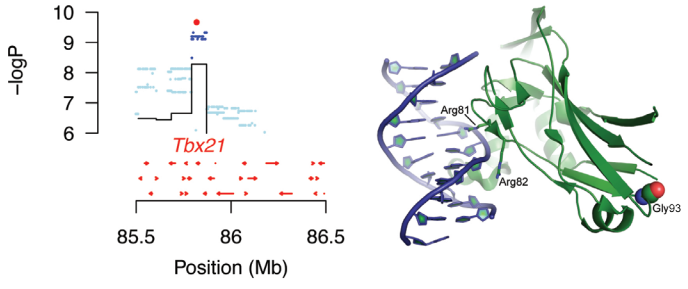
b Heart weight (chromosome 1)



c CD45RC-CD8+ T cells (chromosome 13)



d Proportion of CD4+ T cells expressing CD25 (chromosome 10): TBX21



e Mean red blood cell volume (chromosome 19): ABCB10

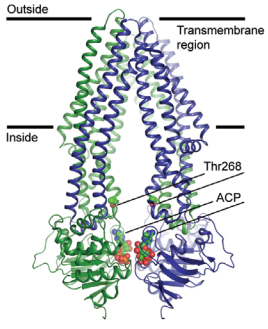


Figure 3 Merge analysis to identify causative genes and sequence variants. The top three panels (a – c) show, on the left, the scans for a whole chromosome, with the name of the phenotype. The black lines represent the haplotype analysis and the blue dots are the merge analysis results of testing for association with all sequence variants identified in the progenitor strains. On the right is an enlargement of the highest peak showing the location of candidate variants and genes. Candidate variants are those whose significance exceeds that of the haplotype analysis (i.e. blue dots are above the highest value of the black line). Genes are shown by red arrows. Panel (d) shows candidate variants on chromosome 10 for the proportion of CD4+ cells with high expression of CD25. The highest variant lies within the TBX21 protein. The crystal structure of human TBX5-DNA complex (PDB code 2X6V) maps the location of the rat TBX21 mutation Gly175Arg to the DNA binding domain. The structure of TBX5 (green) complexed with DNA (blue) is shown in ribbon representation. Gly93 is shown as spheres (C atoms in green, O atoms in red N atoms in blue). Gly93 and corresponding Gly175 (rat) are conserved. Side chains of two arginines that mediate interactions with DNA are shown as sticks. Panel (e) shows a candidate variant in the *Abcb10* gene on chromosome 19 for a locus influencing mean red cell volume. The structure of the homodimeric ABCB10 (PDB code 4AYT) is shown in ribbon representation, with the monomers in blue and green. Two ATP analogues (ACP) and side chains of Thr268 are shown as spheres (C atoms in green, O atoms in red N atoms in blue and P in orange). The rat ABCB10 mutation Thr233Met lies in the central cavity of the translocation pathway. Amino acid sequence identity between rat and human ABCB10 is 84% (587 aligned residues); Thr268 in the human protein corresponds to conserved Thr233 in the rat.

At 43 out of 91 non-synonymous candidate variants, where similar protein structures were available [17] we predicted the structural consequences of mutations (for a further 48 candidate variants there were no homologies with known protein structures). Nine genes, listed in Table 3, contained candidate variants for which structural evidence suggests protein structure or interactions might be altered. An example is shown in Figure 3d, for the protein TBX21, encoded by a gene under a QTL influencing the proportion of CD4+ cells with high expression of CD25.) Here the candidate variant changes glycine to arginine (p.Gly175Arg). The additional arginine could alter the DNA-binding characteristics of this protein. Figure 3e shows the crystal structure of human ABCB10, a mitochondrial transporter induced by GATA1 during erythroid differentiation [18, 19]. The candidate variant (p.Thr233Met), predicted to influence mean red cell volume, maps to a position in the protein structure where the residue side chain points to the centre of the transporter channel (Figure 3e). Threonine has a polar uncharged side chain while methionine has a hydrophobic side chain, a difference likely altering transporter function. Third, merge analysis eliminates candidate genes at a QTL that are distant from any candidate variant. This approach confirmed a well-established relationship between a cluster of apolipoprotein genes at a QTL on chromosome 1 and cholesterol biosynthesis (HDL, LDL and total cholesterol). Similarly, merge analysis identified a locus influencing platelet aggregation, on chromosome 4, that harbors the von Willebrand factor gene (*Vwf*), encoding a key glycoprotein involved in blood coagulation. Merge analysis also contributes to an understanding of the pathogenesis of experimental autoimmune encephalomyelitis (EAE), an autoimmune neuroinflammatory disease

with clinical and pathological similarities to multiple sclerosis (MS) [20]. The MHC class II region on chromosome 20 (*Eae1*) is known to influence EAE susceptibility. However, attempts to identify the responsible gene have had limited success. In this study, the two variants most likely to be causative for the QTL on chromosome 20 (i.e. highest merge $\log_{10}$ *P* value) are a variant in an intron of *Btnl2* and a variant 274 bp upstream of *RT1-Db1*, both in the class II region. The human orthologue of *RT1-Db1, HLA-DRB1*, is associated with multiple sclerosis with risk allele *HLA-DRB1*15:01* [21].

Measure	Chr.	QTL location (Mb)	Gene	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Mean response latency	2	80.23- 84.83	*Ctnnd2*	Catenin delta-2	yes	none	-
Femur neck width	1	156.27- 160.9	*Fchsd2*	FCH and double SH3 domains protein 2	yes	none	-
Distal femur total density	2	152.74- 157.22	*Kcnab1*	Voltage-gated potassium channel subunit beta-1	yes	none	-
Femoral neck total density	5	4.03- 8.22	*Eya1*	Eyes absent homolog 1	yes	none	-
Femur midshaft cortical density	6	38.24- 41.52	*Lpin1*	phosphatidate phosphatase LPIN1	yes	none	-
Femur midshaft total area	2	43.96- 48.57	*Ndufs4*	NADH dehydrogenase [ubiquinone] iron-sulfur protein 4, mitochondrial	yes	none	-
Femur work to failure	8	21.57- 26.17	*Dpy19l1*	protein dpy-19 homolog 1	yes	none	-
Lumbar trabecular area	20	21.1- 25.75	*F1LW02_RAT*	Uncharacterized protein	yes	none	-
Heart weight	1	202.15- 206.63	*Shank2*	SH3 and multiple ankyrin repeat domains protein 2	yes	none	-
Area under glycemia curve over baseline	2	80.5- 85.11	*Ctnnd2*	Catenin delta-2	yes	none	-
Hemoglobin concentration	12	1.62- 5.77	*Insr*	Insulin receptorInsulin receptor subunit alphaInsulin receptor subunit beta	yes	none	-
Mean platelet mass	1	193.98- 197.88	*Dock1*	dedicator of cytokinesis protein 1	yes	none	-

Measure	Chr.	QTL location (Mb)	Gene	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Mean platelet mass	9	52.53- 88.11	*ErbB4*	Receptor tyrosine-protein kinase erbB-4ERBB4 intracellular domain	yes	none	-
Platelet clumps	8	100.57- 104.81	*Clstn2*	Calsyntenin-2	yes	none	-
Platelet count	11	14.47- 18.54	*Hspa8*	Heat shock 70kDa protein 8	yes	none	-
Absolute CD25+CD4+ cells	19	50.71- 54.96	*Galnt2*	polypeptide N-acetylgalactosaminyltransferase 2	yes	none	-
Absolute CD8+ T cells	20	1.00- 8.90	*RT1-Db2*	RT1 class II, locus Db2	yes	none	-
Proportion of B cells in white blood cells	10	27.1- 31.59	*D3ZTU5_RAT*	Uncharacterized protein	yes	none	-
Proportion of B cells in white blood cells	20	1.00- 2.66	*Olr1687*	olfactory receptor Olr1687	yes	none	-
Proportion of CD4+ cells expressing CD45RC	13	36.86- 62.54	*Ptprc*	Receptor-type tyrosine-protein phosphatase C	yes	none	-
Proportion of CD4+ cells in T cells	20	14.83- 19.43	*RGD1559903*	Uncharacterized protein	yes	none	-
Proportion of CD8+ cells expressing of CD45RC	13	50.49- 55.97	*Ptprc*	Receptor-type tyrosine-protein phosphatase C	yes	none	-
Proportion of CD8+ cells with high expression of CD25	19	52.29- 56.8	*Sipa1l2*	Signal-induced proliferation-associated 1-like protein 2	yes	none	-
Lowest weight	3	121.45- 126.25	*Pak7*	serine/ threonine-protein kinase PAK 7	yes	none	-
Weight loss compared to day 0	2	169.79- 174.4	*Fam198b*	Protein FAM198B	yes	none	-
Serum alkaline phosphatase	3	18.49- 23.11	*Lrp1b*	low density lipoprotein-related protein 1B (deleted in tumors)	yes	none	-
Serum chloride concentration	9	32.72- 36.5	*Uggt1*	UDP-glucose:glycoprotein glucosyltransferase 1	yes	none	-
Serum triglycerides	4	74.8- 79.28	*Dfna5*	Deafness, autosomal dominant 5	yes	none	-

Measure	Chr.	QTL location (Mb)	Gene	Gene description	Only gene with candidate variants in QTL	Amino acid change with potential effect	Location of the residue, potential effect
Weight loss compared to day 0	20	2.48- 7.07	*RT1-Da*	RT1 class II histocompatibility antigen Da chain	no	p.Thr182Ala	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48- 7.07	*RT1-Da*	RT1 class II histocompatibility antigen Da chain	no	p.Thr182Met	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48- 7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1 beta chain	no	p.His200Arg	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48- 7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1 beta chain	no	p.Thr165Met	Surface exposed, disturbed intermolecular interactions
Weight loss compared to day 0	20	2.48- 7.07	*RT1-Bb*	RT1 class II histocompatibility antigen, B-1 beta chain	no	p.Gln162Arg	Surface exposed, disturbed intermolecular interactions
Expression on RT1B on B cells	17	26.63- 27.55	*Tbc1d7*	TBC1 domain family member 7	no	p.Ser116Leu	Surface exposed, disturbed intermolecular interactions
Proportion of B cells in white blood cells	1	182.36- 186.67	*Itgal*	Integrin alpha L	no	p.Asn891Ser	Abolish glycosylation
Proportion of CD4+ cells with high expression of CD25	10	84.27- 87.32	*Tbx21*	T-box transcription factor TBX21	no	p.Gly175Arg	Surface exposed, additional DNA interactions
Ratio of T cells to B cells	1	183.58- 187.41	*Rabep2*	Rab GTPase-binding effector protein 2	no	p.Ile336Thr	Partially buried, disturbed oligomerization
Ratio of T cells to B cells	1	183.58- 187.41	*Itgal*	Integrin alpha L	no	p.Leu806Ser	Surface exposed,disturbed intermolecular interactions
Mean corpuscular red cell volume	19	53.11- 55.80	*Abcb10*	ATP-binding cassette, sub-family B (MDR/TAP), member 10	no	p.Thr233Met	Transport channel-exposed, altered transport
Platelet count	12	1.00- 7.47	*Rfc3*	Replication factor C (Activator 1)	no	p.Pro173Ala	Surface exposed, alteration of the alpha helix
Proportion of monocytes in white blood cells	1	250.37- 254.00	*Pdcd11*	Protein RRP5 homolog	no	p.Glu160Gly	Surface exposed

Table 3: Summary of genes identified at QTLs and potential functional variants. The table shows the phenotype measured, the chromosome (Chr.), the start and stop coordinates of the QTL (in megabases Mb), gene symbol and description, whether the gene is the only one at a QTL with candidate variants, whether a variant alters an amino acid and if so the residue changed and potential consequences.

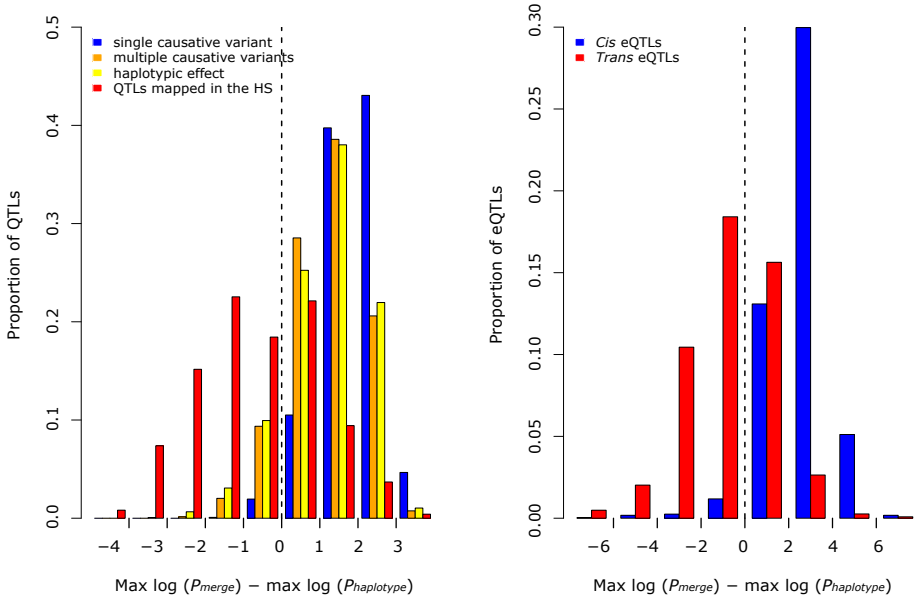


Figure 4 Simulation of causal variants. (a) Plotted are the differences between the maximum negative log *P* values for association of imputed variants and the maximum haplotype-based log *P* values for the rat simulated and real QTLs. In cases where there is a single causal variant at a QTL, the log *P* values of some imputed variants will exceed the significance values from the haplotype analysis, such that the mean of the distribution of the differences between these log *P* values will be greater than zero (blue histogram). The distribution observed for the phenotypic QTLs (red histogram) has a mean less than zero. The results of simulating haplotypic effects are shown in yellow, and the consequence of simulating multiple causative variants are shown in orange. (b) Plotted is a set of 1,386 cis-acting and 7,464 *trans*-acting eQTLs mapped in a mouse heterogeneous stock. The distribution of the differences in log *P* values for the *cis* eQTLs resembles that seen when simulating single causative variants. The distribution for the trans eQTLs is most similar to that for the phenotypic QTLs.

Single variants rarely account for NIH-HS QTL genetic effects

Unexpectedly 212 QTLs (62%) had no candidate variant (Figure 4a). We considered four explanations for this observation: (i) causative variants were missing from the sequence catalogue; (ii) haplotype mapping is biased towards QTLs without candidate variants; (iii) the merge analysis underestimated statistical significance compared to single marker association; (iv) the presence of multiple causal variants. First, causal variants may have been missed because our sequence data are incomplete. Despite linkage disequilibrium extending over a few megabases, not all variants are tagged by a nearby variant with identical strain distribution

pattern (SDP) in the founders. For example, only 50% of the structural variants are tagged by a SNP lying within 1Mb. However, because only a limited set of possible SDPs exist in the HS, we can test whether missing genotypes are responsible for failure to detect candidate variants. We generated SDPs for all diallelic and tri-allelic variants at every locus within the 212 QTLs and tested each by merge analysis, to see how many would have been candidates. Only 44 QTLs had candidate diallelic variants and 165 had diallelic or triallelic variants. Thus if the effects are attributable to a single diallelic variant that we had failed to sequence, then there are still 168 QTLs (49%) without a candidate variant. If the effects are attributable to a di-allelic or tri-allelic variant, the fraction reduces to 14%. However, triallelic SNPs are very uncommon and therefore unlikely to explain the large number of QTLs without candidate variants. Second, haplotype mapping might simply not be powerful enough to detect candidate variants, or be biased towards QTLs without candidate variants. We addressed the first possibility by simulation and show the results in Figure 4a. In each case we report the distribution of the difference d between maximum $\log_{10}P_{merge}$ and $\log_{10}P_{haplotype}$ values, so that if candidate variants exist then $d > 0$. When simulated QTLs arise from single causal variants, merge analysis does indeed identify candidate variants at almost all QTLs placed at random regions of the genome as well as at the QTLs detected. We also considered the performance of the method at QTLs where a single variant is highly likely to be the causal variant, namely at cis-acting expression QTLs [22, 23]. We tested 1,398 eQTLs detected in the hippocampus of HS mice [24]. We found that the merge analysis identified variants that exceed the haplotype-based test at 97% of QTLs (Figure 4b). Interestingly, when we carried out the same analysis on *trans* eQTLs, the distribution of d values was similar to that seen for the rat phenotypic QTLs (Figure 4b). This difference between *cis* and *trans* eQTLs is true across all log P values, indicating that the difference is not due to lower power to detect *trans* eQTLs. Since mapping QTLs using haplotype analysis might bias results towards finding loci without candidates (a winner's curse is likely to operate), we used merge analysis to map QTLs genome-wide. The two methods do not identify the same QTLs (152 are unique to the merge method) but the merge method identified 16% fewer than the haplotype method. Importantly, only 9% of the merge-identified QTLs had no candidate variants (Supplemental figure 3). Consequently haplotype mapping will overestimate the number of QTLs without a candidate variant while merge analysis underestimates it. Therefore our

best estimate of the proportion of QTLs without candidate variants is obtained from combining both methods. From the set of QTLs found by either merge or haplotype mapping we find that 44% of QTLs cannot be explained by single causal variants (instead of 62% when only the haplotype-based QTLs were considered). Thus while a winner's curse does operate in favour of the haplotype analysis, it cannot account for all QTLs without a candidate variant. The third explanation was that the merge analysis under-estimates statistical significance. We compared the performance of the merge analysis with single marker association at genotyped SNPs. Across all phenotypes, the r^2 between the log *P* values was 0.9; agreement was strongest for the most highly associated SNPs. This result indicates that merge analysis performs as well as SNP analysis. Finally, we investigated the extent to which multiple variants at QTLs would account for our findings. We investigated the consequences of a variety of complex QTL architectures by simulation and show the results in Figure 4a. Simulating multiple causal variants, on different haplotypes, reduced the frequency that any single variant exceeded the maximum haplotype log *P* value, although this was still insufficient to mimic the observed distribution (Figure 4a). Simulating irreducible haplotypic effects arising from the reconstructed haplotype mosaics in the HS (rather than from a selection of sequence variants) also led to fewer QTLs with candidate variants (Figure 4a), although again it did not match the proportion observed with the real QTLs. Our simulations suggest that the presence of multiple causal variants at a locus accounts in part for the failure to find candidate causal variants.

Concordance between species

It is often assumed that genetic loci underlying a phenotype identified in one species are homologous to those underlying the same phenotype in another, and that natural variation within these loci will pinpoint the same genes [25-27]. However, there have been no genome-wide tests of the hypothesis for natural variation. Our data allowed us to examine whether genes and QTLs identified in the NIH-HS overlapped those found for the same phenotypes in a mouse HS [10]. In total 38 measures were common to both studies, and were mapped using the same mixed model method. Only one measure, the ratio of CD4+ to CD8+ T-cells, showed overlap (using a FDR of 10% and looking in the 90% QTL confidence interval) but this was not significant (empirical p-value of 0.1). We repeated the analysis using QTLs called at a lower significance threshold (20th percentile of the

extreme value distribution for each measure) and expanding the width of each QTL to 8Mb. Table 4 shows overlap for eight phenotypes, only two of which were significant at an empirical *P* value of 0.05: serum urea concentration and the ratio of CD4+ to CD8+ T-cells. Overall, genetic variants in orthologous genes rarely contribute to the same phenotype in the two populations. To test whether QTL overlap existed within similar pathways we compared the enrichment of KEGG pathways [28]. Only one measure, the proportion of B cells in the white blood cells, showed significant enrichment of a pathway (corrected *P* value < 0.05). Even at a more relaxed significance threshold of 0.05 (non-corrected for multiple testing), only three measures show significant enrichment in the same KEGG pathways.

Discussion

Using 1,407 outbred rats we have mapped 122 phenotypes and identified 355 QTLs at high resolution. We have shown how combining sequence with high resolution mapping data can lead to the immediate identification of candidate genes, and in some cases to the identification of candidate causal variants at many QTL. We highlight two examples here. The locus on chromosome 10 regulating frequency of CD25+ CD4 T cells, and the frequencies of CD4 and CD8 T cells, has previously been shown to control CD4 and CD8 frequencies in a cross between ACI and F344 [29], both represented in the NIH-HS rats. The amino acid substitution at position 175 (p.Gly175Arg) of the TBX21 protein is a very strong causal candidate at this QTL since this domain is important for DNA interactions. *Tbx21* has been implicated in the genetic control of regulatory T cells [30], a subset of T cells with high surface expression of CD25, and might indirectly regulate the frequency of CD4+ and CD8+ T cells via the transcriptional repressor *Sin3a* [31, 32]. We implicated *Abcb10* in red blood cell differentiation. Evidence from mouse knockouts indicates that this gene is essential for erythropoiesis [18, 19, 33]. The p.Thr233Met mutation positions a larger, bulkier residue into a region that is tightly packed in the open-outwards conformation of ABC transporters, potentially interfering with conformational changes which are essential for transport of the substrate. Two noteworthy features of the genetic architecture of complex traits in the rat emerge from this study: (i) the contrast with human GWAS

findings; (ii) about half of QTLs cannot be attributed to a single causal variant. We discuss these points below. Rat and mouse HS experiments differ from human GWAS in two ways. In the rodent GWAS, far fewer subjects are required to detect a significant effect and fewer loci of larger effect explain more of the variance. In rats the median proportion of heritability explained by joint QTLs is 39.1% (mean 42.3%), in mice 32.2% (mean 42.0%). In humans the equivalent figure is often less than 10%. One explanation for these differences is the markedly different allele frequencies: human populations are characterized by a preponderance of rare alleles (minor allele frequency less than 1%); HS populations have a relatively uniform distribution of minor allele frequencies (Figure 1c). However, it is important to realize that the mouse and rat differ in the degree of segregating variation (in the rat NIH-HS there are 7.2M SNP sites, compared to 10.2M in the mouse HS). In the rat there are 2.8M SNPs per HS strain, the corresponding number in the mouse HS is 4.4M. In other words, total sequence variation per se is not a critical determinant of the explanatory power of the QTLs. Furthermore, the heritabilities of the homologous phenotypes in rat NIH-HS and in HS mice are highly correlated (r^2 = 0.6, P = 0.0002) (Supplemental Figure 4), implying that the additional sequence variation in the mouse does not give rise to an increase in heritability. The failure to detect a single candidate variant at half of rat QTLs was surprising. We showed that while reliance on haplotype mapping can underestimate the number of QTLs without candidate variants, after taking this bias into account (by detecting QTLs with merge and haplotype analysis) there is still a large fraction (44%) of QTLs without candidate variants. The contrast between the 44% figure and the 97% that emerged from an analysis of variants at *cis* eQTLs is striking. It is also notable that the findings from *trans* eQTLs are so similar to those of the rat phenotypes (Figure 4) suggesting that *cis* eQTLs are atypical. Our simulations indicate, but have not proven, that multiple causal variants are in part to blame. At present, we can only conclude that single causal variants are not always responsible for the genetic signal. Whether the lack of single causal variants at many loci is a general feature of loci influencing complex traits or not remains to be determined. One simple interpretation of human GWAS is that each locus represents the presence of a single, relatively common, functional variant. Our results indicate that more complex models are required. Such alternative hypotheses exist, in which for example multiple alleles of varying frequency at the same or closely linked loci, contribute to the

signal. Identifying the correct model of genetic action is critical for finding causative variants, since incorrect assumptions about the number and mode of action of genetic variants reduce power and can lead to false positive results [34]. The extent and nature of sequence diversity may be partly responsible for the complex way sequence variation acts at a QTL. It is sometimes hoped that loci found in the rat could be typed and identified in humans, thus providing a cost-efficient way to find medically relevant genes. We observe some examples where the same loci act in different species, the most notable example being for variation in the ratio of CD4+ to CD8+ T-cells: the locus lies within the MHC in rats, humans [35] and mice [36] and its molecular nature in mouse has been identified as a deletion in the promoter of the Class II gene *RT1-Da* [36]. However formal tests for overlap between rat and mouse at the level of the gene or of a pathway yielded little that was statistically significant. Since the amount of sequence variation segregating within the two HS populations is relatively limited, failure to detect shared loci may be due to sampling. Also, the relatively small number of genes found for each phenotype reduces our power to detect pathways. We suspect that currently it is not possible to accurately assess overlap between the two species. This study strengthens the rat's role as a model organism in physiology and disease. Our mapping and sequencing data provide an important resource for addressing many biomedical questions.

Supplemental information

All supplemental files in this chapter can be downloaded from http://www.hubrecht.eu/research/cuppen/hermsen_thesis.html

URLs

Mapping data are available at http://mus.well.ox.ac.uk/gscandb/rat (see Supplementary Note ("Guidelines to explore the genome scans and integrated sequence data") for directions on how to explore the sequence data at each QTL). Variant calls and inaccessible regions are available at http://www.hubrecht.eu/research/cuppen/suppl_data.html.

Accession numbers

Sequence data for the eight HS founders are available from EBI SRA archive under accession ERP001923. The LE/Stm BAC sequences are available in the NCBI Trace Archive (accessions FO181540, FO181541, FO117626, FO181542, FO117624, FO181543, FO117625, FO117627,

FO117628, FO117629, FO117630, FO117631, and FO117632)

Acknowledgements

The funders we would like to acknowledge are: EU 7th Framework Programme (FP7/2007-2013) under grant agreement HEALTH-F4-2010-241504 (EURATRANS); The Wellcome Trust (090532/Z/09/Z, 083573/Z/07/Z, 089269/Z/09/Z) ; The Swedish Research Council [grant number K2008-66X-20776-01-4]; Harald and Greta Jeanssons Foundation; The Swedish Association for Persons with Neurological Disabilities; Åke Wibergs Foundation; Åke Löwnertz Foundation; Karolinska Institutet funds; the EU 6TH Framework EURATools [grant number LSHG-CT-2005-019015]; Bibbi and Nils Jensens Foundation; Söderbergs Foundation; and Knut and Alice Wallenbergs Foundation. The MICINN (ref. PSI2009-10532; and FPI fellowship to C.M-C); the "Fundació La Maratò TV3" (ref. 092630); the DGR (ref. 2009SGR-0051). The British Heart Foundation (BHFRG/07/005/23633). TJA and SA acknowledge funding from the Imperial BHF Centre of Research Excellence. MJ acknowledges support from Prof. Nanna Svartz foundation, The Swedish Rheumatism Association and The King Gustaf V 80th Anniversary Foundation. DG acknowledges support from the Institute of Cardiometabolism and Nutrition (ICAN, ANR-10-IAHU-05). We are grateful to Tadao Serikawa from the Institute of Laboratory Animals, Graduate School of Medicine, Kyoto University, Japan for the LE/Stm BAC clones. The Human Genome Sequencing Center sequence production teams, Baylor College of Medicine produced the Sanger data for 8 sequenced strains used to define the RATDIV SNP genotyping array. See PMID 15057822 for a list of BCM-HGSC sequencing contributors.

Author Contributions

The writing group included A. Baud, R. Hermsen, V.G., D. Gauguier, P.S., T.O., R. Holmdahl, D. Graham, M.W.M., T.F., A.F.-T., N. Hubner, E.C., R.M. and J.F. The phenotyping group included S.C., D. Gauguier, P.S., M.D., J.O., A.D.B., A.G., N.A., A.O.G.-C., M. Jagodic, T.O., M. Johannesson, J.T., U.N., R. Holmdahl, D. Graham, E.B., N. Huynh, W.H.M., M.W.M., A.F.D., D.L.K., T.F., I.A., S.F., N. Hubner, M.O.-P., E.M.-M., R.L.-A., T.C., G.B., E.V.-C., C.M.-C., S.D.-M., A.T. and A.F.-T. The high-density genotyping array design and analysis group included O.H., D.Z., K.S., G.P., A. Bauerfeind, M.-T.B., M.H., Y.-A.L., C.R., H.S., D.A.W., K.C.W., D.M.M., R.A.G., M.L. and N. Hubner. The sequencing group included R. Hermsen, O.H., N.L., G.P.,

P.T., F.P.R., E.d.B., H.H., S.S.A., T.J.A., P.F., D.J.A., T.K., K.S., N. Hubner, V.G. and E.C. The protein structure group included T.M. and E.Y.J. QTL data analysis was performed by A. Baud, J.F., D.E. and R.M. The project was coordinated by A. Baud, R.L.-A., A.F.D., N. Hubner, M. Johannesson, R. Holmdahl, T.O., D. Gauguier, A.F.-T., R.M., E.C. and J.F.

Methods

Sequencing of HS founder genomes

Genome sequencing

DNA libraries for SOLiD sequencing were generated from genomic DNA from samples of the original rats that were used to create the HS population. The libraries were generated using standard protocols (Life Technologies) and had a median insert size of between 109 and 196 bp. All libraries were sequenced with fragment (50 bp) and paired-end (50+35 bp) runs using SOLiD 4 and SOLiD 5500 sequencers to a depth of at least 22x base coverage for each of the eight HS progenitors and for the strain LE/Stm, which was used to estimate error rates in comparison with hand-finished BAC sequence.

Sequence alignment

Sequence reads were mapped against contigs of the Rnor3.4 rat reference genome assembly (reference strain BN) using BWA v0.5.9 [37] with parameters -c –l 25 –k 2 –n 10. Alignments from different libraries of the same HS progenitor were combined into a single BAM file.

Variant calling

Variant calling was performed independently on each strain. SNPs and short indels (<10bp) were called using a modified Samtools [38] pipeline: Only unambiguously mapped reads were used. Sites with coverage below 4 or over 2000 were not used for SNP calling. Read bases with base-quality below 30 were ignored. Duplicate reads starting at the same position and mapped to the same strand as another read were discarded as likely PCR artifacts. Each of the called alleles had to be supported by at least one read where the variant mapped within the seed part of the read (first 25 bases). Non-reference alleles called with fewer than 3 reads

were set to missing. Variable sites with more than 2 alleles within one founder were set to missing. The remaining variants were considered to be homozygous non-reference alleles (frequency of non-reference call > 2/3) or heterozygous alleles (frequency between 1/3 and 2/3) – however, we set to missing the small number of heterozygote calls as these were likely to be artefacts, for example due to unknown duplications. We later attempted to call all the missing genotypes by imputation (see below).
Copy number variants were called using depth-of-coverage approach implemented in DWAC-Seq v. 0.56 (https://github.com/Vityay/DWAC-Seq) using default parameters. Structural variants (SVs) were called using discordant pair mapping implemented in 1-2-3-SV v. 1.0 (https://github.com/Vityay/1-2-3-SV), requiring unambiguous mapping of both paired tags and at least 4 tag pairs per SV event. SV calls from these tools were merged. Prediction of the functional effect of each variant was performed by Variant Effect Predictor tool VEP 2.1 tool [39]. We defined inaccessible regions of the HS rat genomes in a similar way as was done for mouse genomes [4]. A base was considered as accessible if it did not overlap simple, tandem repeats or low complexity sequence (defined by Dust, source: Ensembl release 66; http://www.ensembl.org), was not covered by more than 150 reads, and average mapping quality was at least 40. Nucleotide positions within 15bp of indels were also considered as inaccessible for SNP calling.

False positive and false negative rates

Thirteen BACs from the strain LE/Stm were sequenced using capillary methods, assembled and manually edited, producing a total of 2.1 Mb finished sequence. BAC sequences were aligned using BLAT [40] For each BAC a single contiguous alignment was obtained, which was used to extract single base changes (SNPs) short indels (1-10 bp) and structural variants (100 bp and above). False positive and false negative rates were estimated from 1.9 Mb of genome sequence syntenic between BACs and genome assembly, excluding low quality BAC sequence (as defined by the BAC finishing team) and inaccessible regions (as defined above). False positive and false negative rates within this 1.9Mb were estimated from the discordance between our allele calls and those in the BACs. Low false positive rates were independently confirmed by analysis of a randomly selected subset of 96 SNPs and 96 indels using PCR-based resequencing. Oligonucleotide primers were selected to amplify 300 bp fragments around the candidate polymorphism. When amplification was successful (SNPs:

84, indels: 80), amplicons were sequenced on an Applied Biosystems ABI 3730XL sequencer using Big-Dye terminator and analyzed with Polyphred software manually. For CNVs and SVs 184 variants were selected and PCR primers were designed in such way that the presence or absence (depending on the variation type) of a PCR product could confirm the presence of the variation. After PCR, samples were run on agarose gel and analyzed manually. Of the 184 amplicons, 93 gave a PCR product. Of these 93, a group of 39 variants that were predicted SVs in the NIH-HS founders, were also confirmed by PCR in BN/NHsdMcwi indicating that these are probably assembly errors in the current reference genome (Rn3.4). Of the remaining 54 variants, 53 gave a banding pattern according to our expectance and in one case the predicted variation type was not correctly predicted.

Sequence divergence

Genotypes and genome accessibility data for HS rats (this study) and HS mice [4] were used to characterize patterns of nucleotide diversity in these two panels. We partitioned each genome into non-overlapping windows such that each window contained 100 kb of accessible sequence (defined relative to the rat BN strain or mouse C57BL6 strain). The number of sequence differences per window was calculated for all windows and for all possible pairs of strains.

Low diversity regions

We found the spatial distribution of pairwise differences in the rat progenitors was bimodal with modes at 0 and 150 SNPs per 100 kb window (Figure 1b). Based on this distribution we defined a region of low nucleotide diversity between two strains as consecutive windows with nucleotide diversity below 13 SNPs per 100kb window.

Phenotyping

Animals

The rat NIH-HS originates from a colony established in the 1980s in NIH [3]. Since its creation, the stock has been bred using a rotational outbreeding regime in order to minimize the extent of inbreeding, drift, and fixation. A full description of the phenotyping protocol is given in Supplementary Note ("Phenotyping"). All procedures were carried out in accordance with

the Spanish legislation on "Protection of Animals Used for Experimental and Other Scientific Purposes" and the European Communities Council Directive (86/609/EEC) on this subject. The experimental protocol was approved by the Autonomous University of Barcelona Ethics committee (permit CEEAH 697).

Quality control, covariate analysis and normalisation of phenotypes
The phenotype data were uploaded to a database (Integrated Genotyping System [41]) in batches over the three years of data collection. All relevant covariates were evaluated for their effect on each measure. The final set of covariates and transformations applied to each phenotype, as well as the number of data points for each measure, are given in Supplementary Table 1.

Genotyping
The RATDIV array was developed as a general SNP genotyping array, applicable both to the rat HS project and other populations of laboratory rats. Full descriptions of the development of the rat array and of the selection of the 265,551 SNPs used in this study are given in the Supplementary Note (sections "Development of the rat array" and "Selection of SNPs for this study" respectively).

Linkage disequilibrium analysis
Linkage disequilibrium (LD) between SNPs in the rat and mouse HS was calculated using PLINK [42] from the genotypes called for the 261K autosomal rat SNPs and 12K autosomal mouse SNPs [10]. In the rat HS, eight regions with very high interchromosomal LD were identified, and excluded from subsequent analyses (Supplemental Table 2). Using UCSC liftover tool [43], these regions mapped in the new rat reference genome assembly (RGSC 5.0) to the regions with which they were in high LD in the current assembly (Rnor3.4).

QTL Mapping

Reconstruction of HS rat genomes as mosaics of founder haplotypes
All genetic analysis was performed using R [44]. We used the R HAPPY package [14] to calculate the descent probabilities from the eight HS founders for each animal at each of 265,551 inter-markers intervals, and then averaged these probabilities over 90kb windows, so that we eventually worked with 24,196 probability matrices. The density of the 265k SNPs was much greater than the density of recombinants in the HS, so the averaging did not cause any reduction in mapping resolution (most QTLs are mapped to intervals over a Mb wide, containing over ten 90kb intervals).

Accounting for confounding in the HS
HS rats with different levels of relatedness were used in this study, including siblings, half sibs, cousins, uncles, great-uncles, etc. This unequal genome-wide genetic similarity means that correlations exist in the HS between distant markers. These long-range correlations (as opposed to short-range correlations due to physical linkage) can be responsible for false associations if not accounted for. We used two methods to control for unequal relatedness: Resample Model Averaging (as implemented in BAGPHENOTYPE [13]) for non normally distributed phenotypes, and Mixed Models for normally distributed phenotypes. Information on the scope and the performance of the methods is given in the Supplementary Note (section "Comparison between mixed models and resample model averaging"). Because most of the phenotypes were normally distributed and the merge analysis was run in the mixed model framework, we present the mixed models briefly here. They were implemented in R so that haplotype mapping could be carried out using the descent probabilities output by HAPPY [14]. The model used to test for association between the ancestral haplotypes segregating at a locus L and phenotypic variation was:

$$y_i = \sum_c \beta_c x_{ic} + \sum_s P_{Li}(s) T_{Ls} + u_i + \epsilon_i \qquad (1)$$

where y_i is the phenotypic value of the rat i, β_c the regression coefficient of covariate c and x_{ic} (the value of the covariate c in rat i). Notably, the covariates include a dummy intercept term. T_{Ls} is the deviation in phenotypic value that results from carrying one copy of a haplotype from strain s at locus L and $P_{Li}(s)$ is the expected number of haplotypes of type s carried by rat i at locus L output by HAPPY [14]. u_i and ε_i are random effects, with $\mathrm{cov}(u_i,u_j) = \sigma_g^2 K_{i,j}$ and $\mathrm{cov}(\varepsilon_i,\varepsilon_j) = \sigma_e^2 I_{i,j}$ where σ_g^2 and σ_e^2 are estimated in the null model (no locus effect, $T_{Ls}=0$) using the R package EMMA [12]. K is the genetic covariance matrix, and is estimated from the genome-wide genotypic data using identity by state (IBS, the proportion of shared alleles between any two animals). The IBS matrix was calculated using the R package EMMA [12]. I is the identity matrix. The total covariance matrix $V = \sigma_g^2 K + \sigma_e^2 I$ can be factorized as $V = A^2$. Writing equation (1) in matrix form, gives

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{P_L T_L} + \boldsymbol{u} + \boldsymbol{\epsilon} \qquad (2)$$

Pre-multiplying (2) by A^{-1} gives a transformed equation

$$(A^{-1}\boldsymbol{y}) = (A^{-1}X)\boldsymbol{\beta} + (A^{-1}P_L)T_L + A^{-1}(\boldsymbol{u} + \boldsymbol{\epsilon}),$$

in which the variance-covariance structure of the random term $A^{-1}(u + \varepsilon)$ is now proportional to a diagonal matrix and so can be fitted as a standard linear model.

Thresholds and confidence intervals

Calculations of the significance thresholds (when the phenotype was analysed with mixed models), inclusion probability thresholds (when it was analysed by resample model averaging), and confidence intervals are described in the Supplementary Note (section "QTL calling").

Incorporation of sequence into QTL mapping

Implementation of the merge analysis in the mixed model framework

Merge analysis is a form of imputation appropriate to HS-type populations whose genomes are mosaics of known haplotypes. Merge analysis asks two questions at each imputed variant: is the variant associated with the

phenotype? (a standard test of association), and is it as significant as the test of haplotype association in the locality of the variant? We implemented merge analysis [6] in a mixed-model framework by comparing model (2)

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{P_L T_L} + \boldsymbol{u} + \boldsymbol{\epsilon}$$

and

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{M_V U_V} + \boldsymbol{u} + \boldsymbol{\epsilon} \quad (3)$$

where V is a sequence variant in interval L, and M_V is the merge matrix for the variant, formed by summing those columns of P_L that carry the same allele at V (each column of P_L represents one founder strain). This can be computed efficiently by defining a matrix B_V that encodes the columns to be merged such that $M_V = P_V B_V$. This test is applied at every variable site in the catalogue of single nucleotide variants that segregate between the 8 HS founders. From a statistical point of view, there is no difference between two variants with the same strain distribution pattern at a locus; they will give the same merge analysis result. Because the two models (2), (3) are nested, the best possible fit (in terms of variance explained) is obtained with the haplotype model (2). If the QTL arises from variation at a single variant V, the fit of the merge model (3) for variant V will be as good as the fit of (2) and its significance will be greater due to the fewer number of degrees of freedom (for a diallelic variant, the degrees of freedom is 1 rather than 7 for the haplotype model). The merge model is fitted by multiplying by A^{-1} as before.

Simulating all possible strain distribution patterns at a QTL

For each QTL lacking variants with a merge log P exceeding the haplotype logP, we looked for unobserved causal variants that might not have been sequenced. We simulated candidate variants with every possible strain distribution pattern (SDPs; 127 possible SDPs for diallelic variants, 1,094 possible SDPs when allowing for 3 alleles). The simulated variants were repeated within each QTL interval.

Simulating different QTL architectures

To investigate the hypothesis that failure to detect candidate variants by

merge analysis reflected a complex architecture of the QTLs, we simulated QTLs arising from a single causal variant, QTLs arising from multiple causal variants within the same locus and/or multiple causal variants at linked loci, and QTLs arising from haplotypic effects not reducible to individual variants. In all cases the phenotypes were simulated from three components: a genetic random effect explaining 20% of phenotypic variation, uncorrelated errors explaining 75% of phenotypic variation, and a single QTL explaining 5% of phenotypic variation. When multiple causal variants were simulated, each explained the same proportion of phenotypic variation (5% divided by the number of causal variants). The effect sizes calculated *a posteriori* could be quite different from their target values due to correlations between the different components of the simulated phenotypes. For the simulations reported in Figure 4a, either a single causal variant was simulated, or nine causal variants in three linked loci (each locus within 2Mb of the central locus and distant by at least 200kb from each other locus). Alternatively, the probabilities P_L were used to simulate irreducible QTLs. We analysed each simulation by merge analysis and when log ($P_{haplotype}$) was between 4 and 6 (to have a similar distribution of log *P* values to that of the rat QTLs) we calculated d = max log (P_{merge}) – max log ($P_{haplotype}$). We compared the distribution of d from the different simulation sets to determine the likely genetic architecture of the QTLs.

eQTL mapping and merge analysis in the mouse Heterogeneous Stock

Hippocampus expression levels in 460 HS mice measured using 12 thousand probes of Illumina Mouse WG-6 v1 BeadArrays [24] were mapped to the mouse ancestral haplotypes in the mixed model framework. QTLs were called in the same way as for the rat QTLs but using a confidence interval of 8Mb and a significance threshold of 4. *cis* eQTLs were defined as within 2Mb of the beginning of the probe, and *trans* eQTLs as those QTLs on a different chromosome from that of the probe or more than 10Mb away from it. Merge analysis was carried out at each eQTL and the difference between the maximum merge logP and the maximum haplotype logP was calculated.

Homology modeling

To assess the potential effect of mutations on protein structure, homology models of target proteins were constructed and analysed. Amino acid

sequences of target proteins were retrieved from Ensembl or UniProt databases [45] and analysed using HHPred [46] web server to identify structures with similar amino acid sequences in the Protein Data Bank18 for homology modelling with MODELLER [47]. Potential location of the mutation-bearing side chains (buried or surface exposed) and effect on the structure-function (e.g. disturbed hydrophobic core) was evaluated manually in PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).

Genetic architecture

Heritability

Heritability is defined as the ratio of the genetic variance component by the sum of the variance components estimated in the null mixed model (covariates but no QTL).

QTL Effect sizes and joint effect sizes

Effect sizes are defined as the ratio between the fitted sum of squares and the total sum of squares in a model with covariates and without genetic random component. Joint effect sizes are defined as the ratio between the fitted sum of squares and the total sum of squares in a model without genetic random component including covariates and all the QTLs called for a given phenotype. Including the genetic random component would underestimate most of the effect sizes because part of the variance would have been attributed to it. Thus the QTL effect sizes reported are probably overestimates.

Number of genes under a QTL

The number of genes under each QTL confidence interval was calculated using Ensembl protein coding genes and genes coding for micro RNAs (downloaded from BioMart [48]).

Overlap with RGD QTLs and with QTLs detected in the mouse HS

The calculation of the overlap between RGD and rat HS QTLs, and between mouse and rat HS QTLs is given in the Supplementary Note (section "Overlap between sets of QTLs").

Pathway analysis for the QTLs detected in the rat and mouse heterogeneous stocks

Kyoto Encyclopedia of Genes and Genomes pathways were retrieved using R KEGG.db package. We used INRICH [49] to find enrichment of pathways in the mouse and rat phenotypic QTLs (as defined by the 90% confidence interval) called at a low significance threshold (20th percentile of the extreme value distribution). We report the empirical and corrected *P* values reported by INRICH.

References

1. Donnelly P: Progress and challenges in genome-wide association studies in humans. Nature 2008, 456(7223):728-731.
2. Flint J, Mackay TF: Genetic architecture of quantitative traits in mice, flies, and humans. Genome Res 2009, 19(5):723-733.
3. Hansen C, Spuhler K: Development of the National Institues of Health genetically heterogeneous rat stock. Alcoholism: Clinical and Experimental Research 1984, 8(5):477-479.
4. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M et al: Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 2011, 477(7364):289-294.
5. Talbot CJ, Nicod A, Cherny SS, Fulker DW, Collins AC, Flint J: High-resolution mapping of quantitative trait loci in outbred mice. Nature Genetics 1999, 21(3):305-308.
6. Yalcin B, Flint J, Mott R: Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. Genetics 2005, 171(2):673-681.
7. Mayosi BM, Keavney B, Watkins H, Farrall M: Measured haplotype analysis of the aldosterone synthase gene and heart size. European journal of human genetics : EJHG 2003, 11(5):395-401.
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(6982):493-521.
9. Serikawa T, Mashimo T, Takizawa A, Okajima R, Maedomari N, Kumafuji K, Tagami F, Neoda Y, Otsuki M, Nakanishi S et al: National BioResource Project-Rat and related activities. Experimental animals / Japanese Association for Laboratory Animal Science 2009, 58(4):333-341.
10. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, Taylor MS, Rawlins JN, Mott R, Flint J: Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat Genet 2006, 38(8):879-887.
11. Johannesson M, Lopez-Aumatell R, Stridh P, Diez M, Tuncel J, Blazquez G, Martinez-Membrives E, Canete T, Vicens-Costa E, Graham D et al: A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: the NIH heterogeneous stock. Genome Res 2009, 19(1):150-158.
12. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: Efficient control of population structure in model organism association mapping. Genetics 2008, 178(3):1709-1723.
13. Valdar W, Holmes CC, Mott R, Flint J: Mapping in structured populations by resample model averaging. Genetics 2009, 182(4):1263-1277.
14. Mott R, Talbot CJ, Turri MG, Collins AC, Flint J: A method for fine mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci U S A 2000, 97(23):12649-12654.
15. Israely I, Costa RM, Xie CW, Silva AJ, Kosik KS, Liu X: Deletion of the neuron-specific protein delta-catenin leads to severe cognitive and synaptic dysfunction. Current biology : CB 2004, 14(18):1657-1663.
16. Berkel S, Marshall CR, Weiss B, Howe J, Roeth R, Moog U, Endris V, Roberts W, Szatmari P, Pinto D et al: Mutations in the SHANK2 synaptic scaffolding gene in autism spectrum disorder and mental retardation. Nat Genet 2010, 42(6):489-491.
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. Nucleic acids research 2000, 28(1):235-242.
18. Shirihai OS, Gregory T, Yu C, Orkin SH, Weiss MJ: ABC-me: a novel mitochondrial transporter induced by GATA-1 during erythroid differentiation. The EMBO journal 2000, 19(11):2492-2502.

19. Hyde BB, Liesa M, Elorza AA, Qiu W, Haigh SE, Richey L, Mikkola HK, Schlaeger TM, Shirihai OS: The mitochondrial transporter ABC-me (ABCB10), a downstream target of GATA-1, is essential for erythropoiesis in vivo. Cell death and differentiation 2012, 19(7):1117-1126.

20. Wallström EOT: Rat Models of Experimental Autoimmune Encephalomyelitis. In: Sourcebook of Models for biomedical research. Edited by Conn PJ. Otowa: Humana Press Inc; 2007.

21. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE et al: Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 2011, 476(7359):214-219.

22. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE et al: DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 2012, 482(7385):390-394.

23. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK: High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 2008, 4(10):e1000214.

24. Huang GJ, Shifman S, Valdar W, Johannesson M, Yalcin B, Taylor MS, Taylor JM, Mott R, Flint J: High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. Genome Res 2009, 19(6):1133-1140.

25. Jagodic M, Colacios C, Nohra R, Dejean AS, Beyeen AD, Khademi M, Casemayou A, Lamouroux L, Duthoit C, Papapietro O et al: A role for VAV1 in experimental autoimmune encephalomyelitis and multiple sclerosis. Science translational medicine 2009, 1(10):10ra21.

26. Swanberg M, Lidman O, Padyukov L, Eriksson P, Akesson E, Jagodic M, Lobell A, Khademi M, Borjesson O, Lindgren CM et al: MHC2TA is associated with differential MHC molecule expression and susceptibility to rheumatoid arthritis, multiple sclerosis and myocardial infarction. Nat Genet 2005, 37(5):486-494.

27. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, Bakker SF, Bardella MT, Bhaw-Rosun L, Castillejo G et al: Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nat Genet 2011, 43(12):1193-1201.

28. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research 1999, 27(1):29-34.

29. Brenner M, Laragione T, Yarlett NC, Gulko PS: Genetic regulation of T regulatory, CD4, and CD8 cell numbers by the arthritis severity loci Cia5a, Cia5d, and the MHC/Cia1 in the rat. Mol Med 2007, 13(5-6):277-287.

30. Koch MA, Tucker-Heard G, Perdue NR, Killebrew JR, Urdahl KB, Campbell DJ: The transcription factor T-bet controls regulatory T cell homeostasis and function during type 1 inflammation. Nature immunology 2009, 10(6):595-602.

31. Chang S, Collins PL, Aune TM: T-bet dependent removal of Sin3A-histone deacetylase complexes at the Ifng locus drives Th1 differentiation. J Immunol 2008, 181(12):8372-8381.

32. Cowley SM, Iritani BM, Mendrysa SM, Xu T, Cheng PF, Yada J, Liggitt HD, Eisenman RN: The mSin3A chromatin-modifying complex is essential for embryogenesis and T-cell development. Molecular and cellular biology 2005, 25(16):6990-7004.

33. Liesa M, Luptak I, Qin F, Hyde BB, Sahin E, Siwik DA, Zhu Z, Pimentel DR, Xu XJ, Ruderman NB et al: Mitochondrial transporter ATP binding cassette mitochondrial erythroid is a novel gene required for cardiac recovery after ischemia/reperfusion. Circulation 2011, 124(7):806-813.

34. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT et al: Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 2010, 465(7298):627-631.

35. Ferreira MA, Mangino M, Brumme CJ, Zhao ZZ, Medland SE, Wright MJ, Nyholt DR, Gordon S, Campbell M, McEvoy BP et al: Quantitative trait loci for CD4:CD8 lymphocyte ratio are associated with risk of type 1 diabetes and HIV-1 immune control. Am J Hum Genet 2010, 86(1):88-92.

36. Yalcin B, Nicod J, Bhomra A, Davidson S, Cleak J, Farinelli L, Osteras M, Whitley A, Yuan W, Gan X et al: Commercially available outbred mice for genome-wide association studies. PLoS Genet 2010, 6(9).

37. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25(14):1754-1760.

38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25(16):2078-2079.

39. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F: Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010, 26(16):2069-2070.

40. Kent WJ: BLAT--the BLAST-like alignment tool. Genome Res 2002, 12(4):656-664.

41. Fiddy S, Cattermole D, Xie D, Duan XY, Mott R: An integrated system for genetic analysis. BMC Bioinformatics 2006, 7:210.

42. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007, 81(3):559-575.

43. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F et al: The UCSC Genome Browser Database: update 2006. Nucleic acids research 2006, 34(Database issue):D590-598.

44. R-Development-Core-Team: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2004.

45. Magrane M, Consortium U: UniProt Knowledgebase: a hub of integrated protein data. Database : the journal of biological databases and curation 2011, 2011:bar009.

46. Soding J, Biegert A, Lupas AN: The HHpred interactive server for protein homology detection and structure prediction. Nucleic acids research 2005, 33(Web Server issue):W244-248.

47. Eswar N, Eramian D, Webb B, Shen MY, Sali A: Protein structure modeling with MODELLER. Methods Mol Biol 2008, 426:145-159.

48. Kasprzyk A: BioMart: driving a paradigm change in biological data management. Database : the journal of biological databases and curation 2011, 2011:bar049.

49. Lee PH, O'Dushlaine C, Thomas B, Purcell SM: INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics 2012, 28(13):1797-1799.

4

Chapter 4

Multilevel effects of non-coding genetic variation on regulatory elements and chromatin organization

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

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

*These authors contributed equally to this work

Abstract

Disease-linked variants identified by genome-wide association studies (GWAS) are often located outside protein-coding genes, but the effects of these variants are poorly understood. To assess the effects of non-coding variants on gene expression, we analyzed the interactions among genetic variation, epigenetic regulatory elements, three-dimensional (3D) chromatin organization, and transcription in ten inbred rat strains. We found higher diversity across strains for enhancers than promoters and part of this variability was linked to nucleotide or structural variants inside these elements. Genetic variants in enhancers and promoters also resulted in transcription level differences. The 3D chromatin organization was found to be largely conserved among strains, although regional exceptions were identified. We show that differences in 3D organization were associated with increased genetic variation, including structural variations. Interestingly, differences in 3D chromatin organization corresponded more strongly with changes in regulatory elements than with differentially expressed genes. Our findings demonstrate that genetic variation influenced gene expression through multiple regulatory mechanisms. Therefore multilevel analyses are required to properly understand the functional effects of non-coding genomic variation.

Introduction

The majority (93%) of disease-associated variants that are detected by human genome-wide association studies (GWAS) reside in the non-coding part of the genome [1-3]. This is not completely unexpected, because only ~2–3% of the human genome encodes proteins [4, 5], and a much larger proportion of the genome has been shown to have different biological functions, which are often related to regulatory mechanisms [6-8]. However, the effects of genetic variation in non-coding genomic regions are still poorly understood [9, 10]. The regulation of gene expression is an intricate process that involves multiple layers of interactions. At the DNA sequence level, there are two main types of regulatory elements: promoters, which are located at the transcriptional start sites (TSS) of genes, and enhancers, which regulate transcription at sites more distally from the TSS. Functional regulatory element activity depends on the recruitment of general or tissue-specific transcription factors (TFs). For example, forkhead box protein A1 (FOXA1), hepatocyte nuclear factor 4 alpha (HNF4A), and CCAAT/enhancer-binding protein alpha (CEBPA) are

tissue-specific TFs that regulate the transcription of liver-specific genes [11]. Enhancers and promoters are marked by epigenetic modifications. Acetylation of the lysine 27 residue of histone H3 (H3K27ac) is associated with both active promoters and active enhancers [12]. Tri-methylation of the lysine 4 residue of histone H3 (H3K4me3) discriminates between the two types of regulatory elements and is only associated with active promoters [13]. In addition to regulatory DNA elements, the 3D folding structure of the genome contributes to gene expression regulation. Genome-wide adaptations of chromatin conformation capture (3C)-based techniques (e.g. Hi-C [14] or chromatin interaction analysis with paired-end tag sequencing [15]) now provide the means to study chromatin organization in great detail. These methods have revealed that the nuclear 3D organization divides chromosomes into large units of transcriptional activity or inactivity, which are referred to as topological domains [16]. Recently, a series of *in vitro* studies [17-20] revealed a direct effect of non-coding single nucleotide variants (SNVs) on TF binding. These studies, as well as all GWAS, primarily focus on the effects of single nucleotide changes. SNVs are the most studied type of genetic variation because of their high detection efficiency. In addition, the effects of this variant type are relatively easy to predict, especially when they reside in protein-coding genome regions [21, 22]. However, another category of genetic variation is structural genome variation, which includes deletions, duplications, translocations, and inversions. Even though structural variants are less common than SNVs, they affect many more bases per genome [23]. Therefore, understanding the functional effects of structural changes in the non-coding genome will be important for the proper interpretation of personal genomes for disease diagnosis.

Here, we systematically studied the effects of genomic variation on various levels of transcription regulation *in vivo*. Liver tissues, which are relatively homogeneous, from ten inbred rat strains were studied. Inbred strains were used to avoid difficulties with the analysis of allele-specific effects at heterozygous positions. We integrated structural and single nucleotide genomic variations from whole genome sequencing, regulatory element (enhancers and promoters) profiles obtained by chromatin immunoprecipitation sequencing (ChIP-seq), transcriptional profiling (RNA-seq), and genome-wide chromatin organization maps obtained by Hi-C to dissect the different layers by which genomes function *in vivo*.

Results

Landscape of regulatory elements shows diversity among strains

We generated a map of the active regulatory elements in liver tissues of ten inbred rat strains (ACI, BN, BUF, F344, M520, MR, WN, WKY, BN-*Lx*, and SHR) by H3K4me3 and H3K27ac ChIP-seq (Fig. 1A and 1B). This resulted in the identification of 52,240 active enhancers and 14,511 active promoters in all strains (Supplemental Table 1 and 2). To test the comprehensiveness of this set of regulatory elements, we analyzed the overlap with previously generated rat liver ChIP data of liver-specific TFs (HNF4A, CEBPA, and FOXA1) [11].

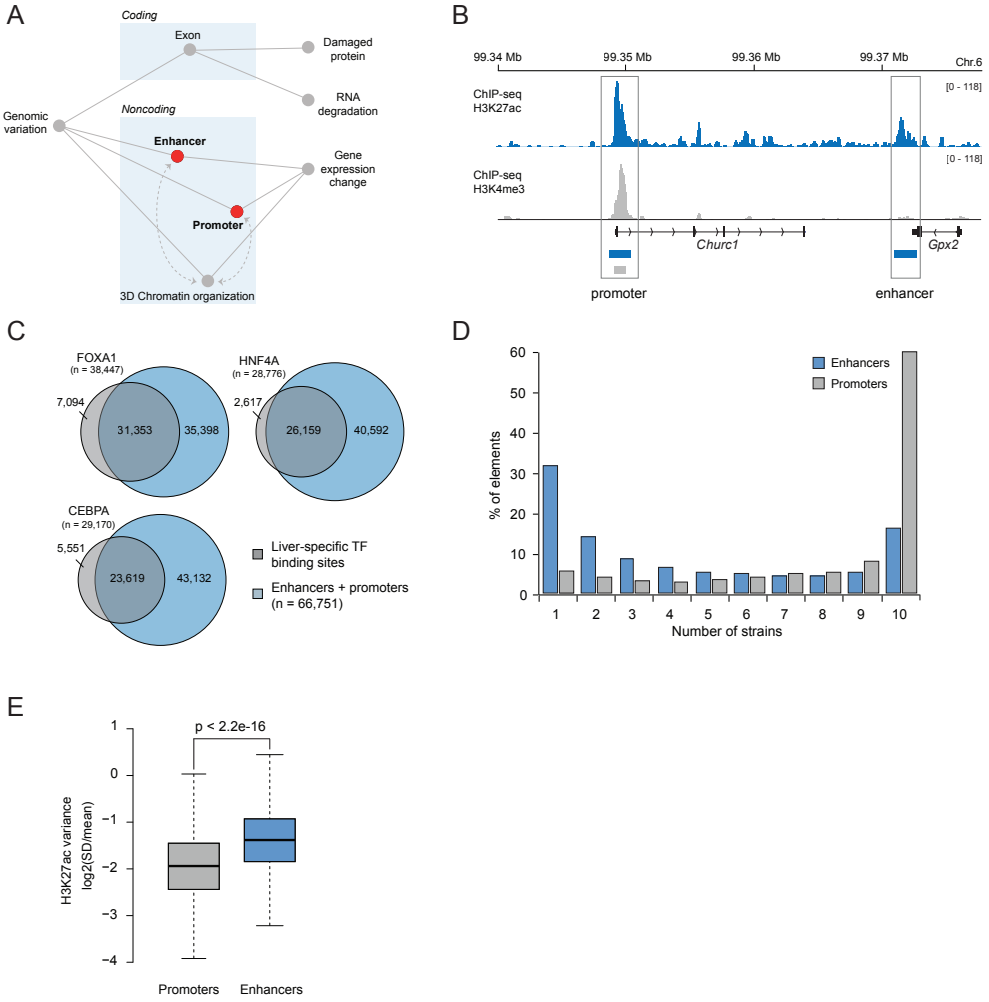


Figure 1. Detection and comparison of promoters and enhancers in liver tissues of ten rat strains. (A) Schematic showing the position of promoters and enhancers in the interaction network. (B) Identification of active enhancers and promoters. Promoters show enrichment for histone H3 lysine 4 tri-methylation (H3K4me3; grey) and histone H3 acetyl lysine 27 (H3K27ac; blue) modifications. Enhancers are solely marked by H3K27ac and lack H3K4me3 enrichment. (C) Venn diagrams displaying the overlap of liver-specific transcription factor (TF)-binding positions of forkhead box protein A1 (FOXA1), hepatocyte nuclear factor 4A (HNF4A), and CCAAT/enhancer-binding protein alpha (CEBPA) with active enhancers and promoters. (D) Examples of liver-specific TF binding localized to gene promoters and enhancers. (E) Bar plot showing the distribution of enhancers and promoters among the ten rat strains. (F) Boxplot showing the relative variance in H3K27ac levels in promoters and enhancers.

Overall, the enhancers and promoters captured 80% of the TF peaks (Fig. 1C). Approximately half of the liver TF-bound enhancers (n = 9,356) bound to all three liver-specific TFs, thus indicating high co-localization between tissue-specific TFs (Supplemental Fig. 1). We also found that 35.9% of all active enhancers and 47.9% of all active promoters were bound by at least one of the three liver TFs. The remainder of the elements likely harbored binding sites for TFs that were not tested in this study. To investigate the amount of variation that is found between strains at the epigenetic level, we first examined the distribution of each regulatory element among strains. The vast majority (~60%) of all promoters were present in all ten livers, whereas ~35% were shared between 2–9 strains, and only 5% were strain specific. The distribution of enhancer elements was more variable. Only 16% of the enhancers were present in all ten strains, 53% were found in 2–9 strains, and 31% were strain specific (Fig. 1D). To substantiate these findings, we determined the relative variance in H3K27ac levels for the promoters and enhancers among all strains. In agreement with the previous results, enhancers showed significantly higher relative variance in H3K27ac levels than promoters ($p < 2.2e-16$, Fig. 1E).

Genetic variation contributes to epigenetic diversity

We recently sequenced the complete genomes of the ten rat strains, so a full inventory of strain-specific and shared genetic variants is available [24, 25]. This inventory allowed us to determine the contribution of genetic variation to the epigenetic diversity that was found among regulatory elements (Fig. 2A). The genome-wide density of SNVs of all ten strains combined was 2.8 SNVs/kb. SNV density was lower in the functional regions of the genome, with 1.7 SNVs/kb in coding regions, 2.6 SNVs/kb in the active enhancers, and 2.2 SNVs/kb in the promoters defined above. The higher level of genetic variation in the enhancers versus promoters corresponded with the larger variation at the epigenetic

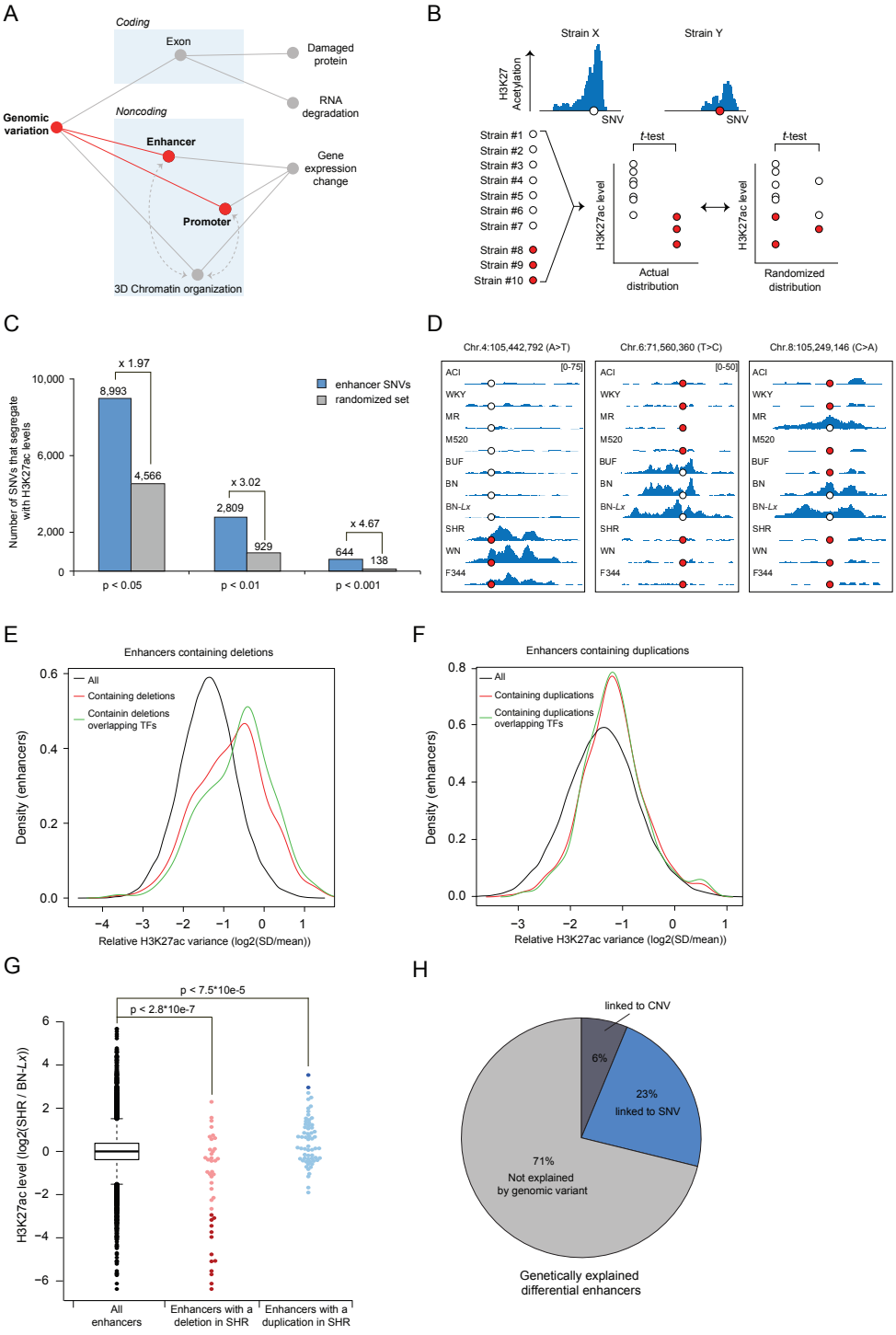

A
Coding
Exon
Damaged protein
RNA degradation
Genomic variation
Noncoding
Enhancer
Gene expression change
Promoter
3D Chromatin organization
B
Strain X
Strain Y
H3K27 Acetylation
SNV
SNV
Strain #1
Strain #2
Strain #3
Strain #4
Strain #5
Strain #6
Strain #7
Strain #8
Strain #9
Strain #10
t-test
t-test
H3K27ac level
H3K27ac level
Actual distribution
Randomized distribution
C
Number of SNVs that segregate with H3K27ac levels
enhancer SNVs
randomized set
x 1.97
8,993
4,566
x 3.02
2,809
929
x 4.67
644
138
p < 0.05
p < 0.01
p < 0.001
D
Chr.4:105,442,792 (A>T)
Chr.6:71,560,360 (T>C)
Chr.8:105,249,146 (C>A)
[0-75]
[0-50]
ACI
WKY
MR
M520
BUF
BN
BN-Lx
SHR
WN
F344
E
Enhancers containing deletions
All
Containing deletions
Containin deletions overlapping TFs
Density (enhancers)
Relative H3K27ac variance (log2(SD/mean))
F
Enhancers containing duplications
All
Containing duplications
Containing duplications overlapping TFs
Density (enhancers)
Relative H3K27ac variance (log2(SD/mean))
G
p < 7.5*10e-5
p < 2.8*10e-7
H3K27ac level (log2(SHR / BN-Lx))
All enhancers
Enhancers with a deletion in SHR
Enhancers with a duplication in SHR
H
linked to CNV
6%
23% linked to SNV
71% Not explained by genomic variant
Genetically explained differential enhancers

Figure 2. Epigenetic diversity is linked to genomic variation. (A) Schematic showing the position of genetics and epigenetics in the interaction network. (B) Schematic representation of the analysis that was 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 the Student's t-test. (C) Bar plot showing the number of SNVs that positively correlated 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. (D) Three examples of SNVs that co-segregate with H3K27ac levels in enhancers without overlapping a liver-specific TF peak. (E and F) Density plots depicting the variance in H3K27ac levels across strains for enhancers targeted by deletions (E) or duplications (F) (red lines). The green line shows the variance in H3K27ac levels for deletions or duplications that targeted a TF-binding site within the enhancer. (G) H3K27ac levels in SHR compared to BN-*Lx* in all enhancers, enhancers with deletions, and enhancers with duplications. Dark red and dark blue points represent differential enhancers (FDR < 0.05; log(average CPM) > 2). (H) Pie chart depicting the percentage of differential enhancers between SHR and BN-Lx (FDR < 0.05; log(average CPM) > 2) that overlap specific genetic variants.

level in the enhancers. Next, we determined the segregation of SNVs with H3K27ac levels to identify potential causal links between specific variants and the observed differences in regulatory elements. We defined the strain distribution pattern for each SNV that was located in an enhancer and selected the SNVs that were present in at least three strains and no more than seven strains (n = 97,030). The difference in H3K27ac levels between alleles was tested, and the significance of the associations was assessed by randomizing the allele distribution (Fig. 2B). A clear enrichment of significant SNV-H3K27ac level associations in the real SNV distribution compared to the randomized sets was observed. This enrichment increased with decreasing p-values for the associations (8,993 versus 4,566 with $p < 0.05$; 644 versus 138 with $p < 0.001$) (Fig. 2C and 2D). Similar enrichments were obtained for SNVs in the promoters (Supplemental Fig. 2A). These data demonstrate that H3K27ac variation in both enhancers and promoters had an important genetic basis. One of the ways by which a genetic variant could affect epigenetic characteristics is by altering a TF-binding site. We tested the frequency of H3K27ac associations in all SNVs versus SNVs that overlapped known TF motifs. Increased associations of H3K27ac levels with several liver TFs, such as CEBPA (Chi-square test, $p < 0.032$) and HNF4A (Chi-square test, $p < 0.0037$), but not FOXA1, were observed (Supplemental Table 3). These results indicate that a subset of the SNVs affect H3K27ac levels by altering TF binding. In addition to SNVs, structural genomic variants may also affect enhancer function. We previously determined copy number variants (CNVs) for these ten strains using the read depth of coverage analysis (DOC) [24, 25]. DOC detects structural copy number changes, including deletions and duplications, by quantitating the differences in the number of whole genome sequencing reads. In total, 1,780 deletions and 1,176 duplications were present in one or more of the ten strains (Supplemental Table 1). Eight hundred

twenty-seven enhancers overlapped with CNVs, and these enhancers had an increased variance in H3K27ac levels across the ten strains (Fig. 2E and 2F). The effect on H3K27ac variation was more profound for deletions than for duplications (means: overall = -1.38, deletions = -8.83 [t-test; p-value 2.2*10-16], and duplications = -1.17 [t-test; p-value 1.8*10-14]). For deletions that overlapped TF sites, the change in H3K27ac levels was even greater (-0.66, t-test; p-value 2.2*10-16) (Fig. 2E). These data demonstrate that CNVs contributed significantly to the epigenetic variation between different genetic backgrounds. To further quantitate the contribution of CNVs and SNVs to enhancer diversity, we directly compared two strains: BN-Lx and SHR. As expected, CNVs decreased and increased H3K27ac levels for deletions and duplications, respectively (Fig. 2G, examples in Supplemental Fig. 2B and 2C). The mean log2(SHR/BN-*Lx*) was -0.007 in all enhancers, -2.25 in enhancers with a deletion in SHR (t-test, p-value < 2.8*10e-7), and 0.56 in enhancers with a duplication in SHR (t-test p-value < 7.5*10e-5). For the quantitation of genetic effects, we selected enhancers that differed the most between strains, based on H3K27ac levels (false discovery rate [FDR] < 0.05; log10(average Counts Per Million) > 2). For this set of 208 differential enhancers, we determined how many were linked to genomic variants. Thirteen and two differential enhancers showed direct overlap with deletions and duplications, respectively (Fig. 2G). Furthermore, 47 enhancers showed a significant correlation between the SNV and H3K27ac variation ($p < 0.05$), as tested in the SNV-H3K27ac analysis that was performed on all ten rats (Fig. 2C). Overall, 62 differential enhancers were associated with genomic variants (Fig. 2H and Supplemental Table 4).

SNVs in regulatory elements are associated with altered gene expression

Next, we examined the functional consequences and effects of genetic and epigenetic variations on gene expression levels (Fig. 3A). RNA-seq was performed on liver tissues of the ten strains to determine which SNVs associate with gene expression. All SNVs that were located within 1 Mb of the TSS were tested for association with the expression level of a given gene (Fig. 3B). The proportion of SNVs that associated with gene expression levels was significantly higher in those that were located in enhancer regions (Fig. 3C). This enrichment was even greater in the group of SNVs that was located in promoters (Fig. 3C). These data demonstrate that genetic variation in regulatory elements was associated with changes in gene expression.

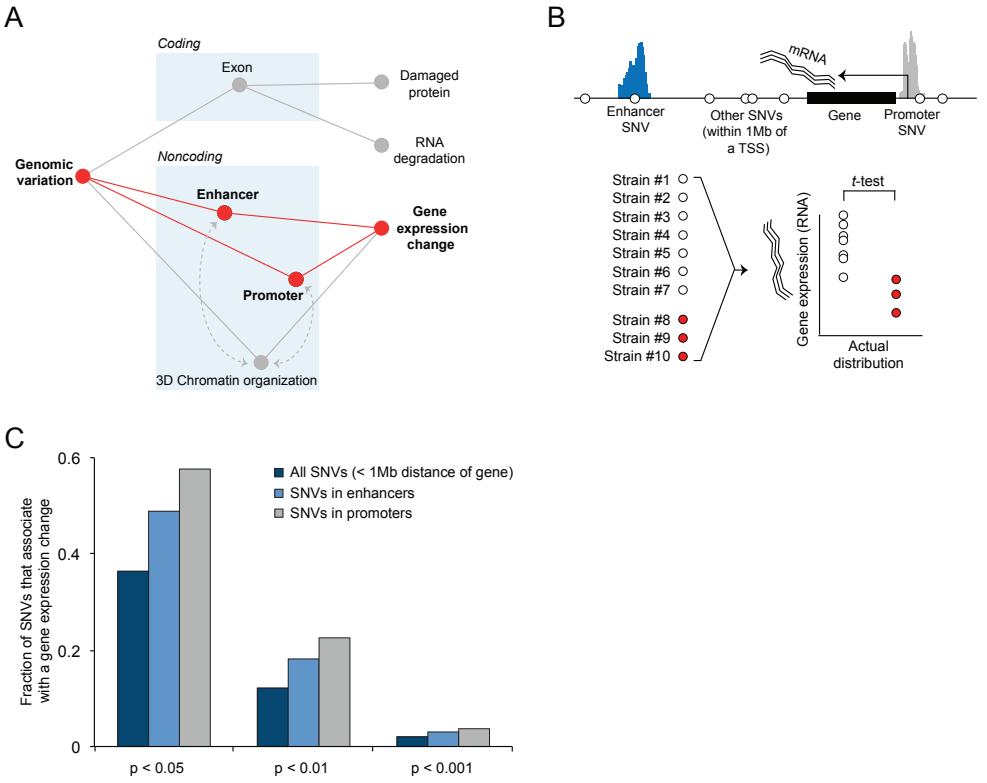


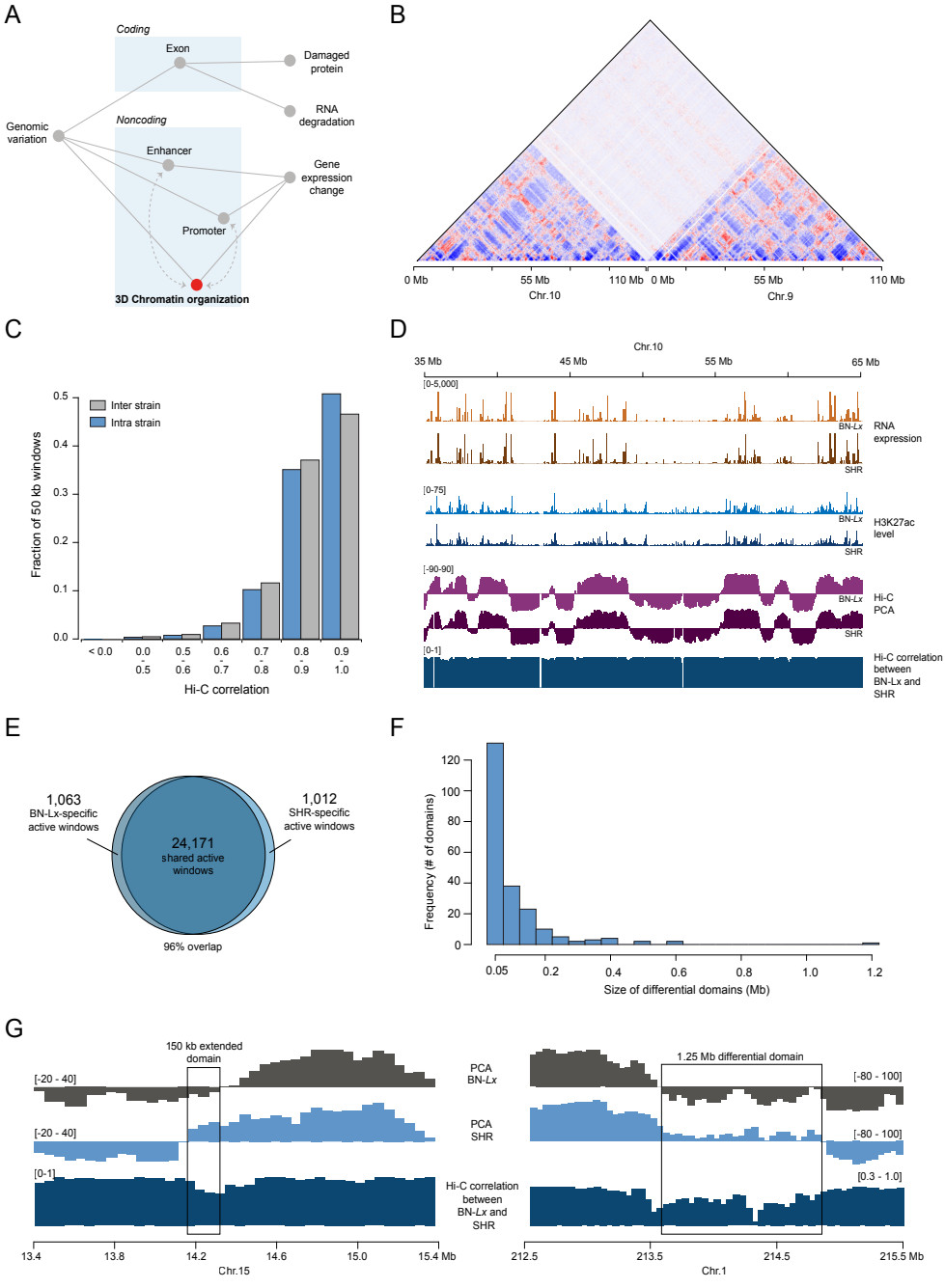
Figure 3. SNVs in regulatory elements are enriched for association with gene expression. (A) Schematic showing the positions of gene expression and genetic variation in the interaction network. (B) Schematic representation of the analysis that was used to test the association of SNVs with gene expression levels. For each gene, RNA levels of strains with and without the SNV were compared using the Student's t-test. All SNVs within 1 Mb were tested. (C) Bar plot showing the number of SNVs that positively correlated with the expression level of at least one gene at p-value cut-offs of 0.05, 0.01, and 0.001.

3D chromatin organization is highly conserved in different genetic backgrounds

On top of a genetic and epigenetic structure, the genome is organized in 3D (Fig. 4A). Hi-C is a method to study 3D DNA organization, and it interrogates spatial interactions between genomic fragments. We performed Hi-C experiments on the liver tissues of BN-*Lx* and SHR, each in triplicate, to investigate the inter-strain variation in the higher-order chromatin organization. Hi-C interaction maps were created by analyzing interaction frequencies per 50-kb window. In line with previous reports [14, 26], the interaction maps show that most interactions were found on the same chromosome (Fig. 4B). Overall, the interaction maps looked very similar when we compared samples from the same strain and between

two different strains. The similarity between samples was quantified by calculating the correlation for each 50-kb genomic window. The correlation between samples of the same strain was slightly higher than the correlation between samples of different strains (Fig. 4C). A correlation > 0.9 was observed in 50.8% and 46.6% of the windows in intra- and inter-strain comparisons, respectively. In general, the correlation between samples was very high, and 85.9% and 83.7% of the windows showed a correlation > 0.8 in intra- and inter-strain comparisons, respectively. This suggests that chromatin folding is largely conserved between different genetic backgrounds, although exceptions are detected. Previous 4C and Hi-C studies have shown that the genome is spatially organized into active and inactive domains [14, 26], which can be dissected by applying the Principal Component Analysis (PCA) [14]. To pinpoint the differences between strains at the highest resolution, we pooled the data of three biological replicates per strain and performed the PCA on these pools at a 50-kb resolution (Fig. 4D). Consecutive 50-kb windows with positive PCA values were joined into domains. We identified 940 active domains in BN-*Lx* and 991 in SHR, with a median size of 350 kb in both strains. Overall, the active domains comprised 1.26 Gb of the BN-*Lx* and the SHR genomes, which equaled approximately half of the rat genome. As expected, the active domains contained the majority of the expressed genes and enhancers (83% and 80%, respectively) (Fig. 4D). The active domains in BN-*Lx* and SHR had a high degree of overlap, in which 96% of the 50-kb windows found in the active domains in BN-*Lx* were also found in the active domains in SHR, and vice versa (Fig. 4E). After combining consecutive 50-kb windows, 596 and 615 genomic regions were solely located in the active domains in SHR and BN-Lx, respectively. To confirm that the differences between strains were not the result of a single outlier, PCA was separately performed on each sample.

Figure 4. Hi-C results showing similarities in the higher-order chromatin structure between BN-*Lx* and SHR. (A) Schematic showing the position of the higher-order chromatin structure in the interaction network. (B) Example of a Hi-C interaction profile between chromosomes 10 and 11. Red and blue depict over- and under-represented interactions, respectively. Very few inter-chromosomal interactions are observed. (C) Correlation scores for 50-kb bins derived from Hi-C comparisons between and within BN-*Lx* and SHR. (D) An example of active and inactive transcriptional domains, as determined by principal component analysis, H3K27ac levels, and RNA-seq expression data for a 30-Mb region of chromosome 10. (E) Overlap between active domains in BN-*Lx* and SHR per 50-kb window. (F) Sizes of the differential domains between BN-*Lx* and SHR. (G) Examples of differential domains.

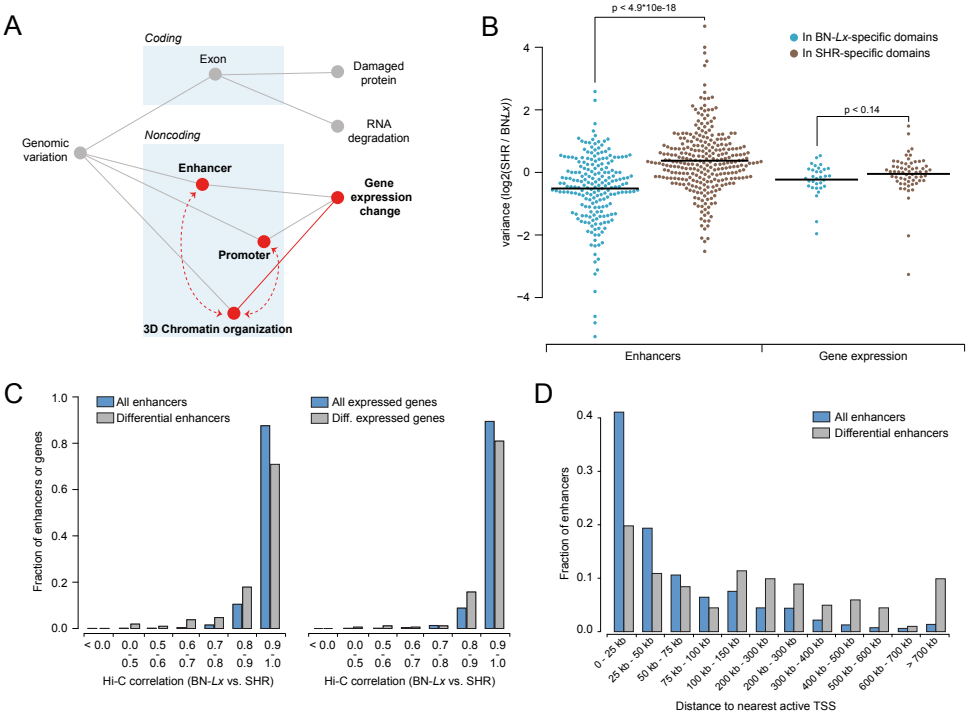

A
Coding
Exon
Damaged protein
RNA degradation
Genomic variation
Noncoding
Enhancer
Gene expression change
Promoter
3D Chromatin organization
B
0 Mb
55 Mb
110 Mb
Chr.10
0 Mb
55 Mb
110 Mb
Chr.9
C
Inter strain
Intra strain
Fraction of 50 kb windows
Hi-C correlation
D
Chr.10
35 Mb
45 Mb
55 Mb
65 Mb
RNA expression
H3K27ac level
Hi-C PCA
Hi-C correlation between BN-Lx and SHR
E
1,063
BN-Lx-specific active windows
24,171
shared active windows
1,012
SHR-specific active windows
96% overlap
F
Frequency (# of domains)
Size of differential domains (Mb)
G
150 kb extended domain
1.25 Mb differential domain
PCA BN-Lx
PCA SHR
Hi-C correlation between BN-Lx and SHR
Chr.15
Chr.1

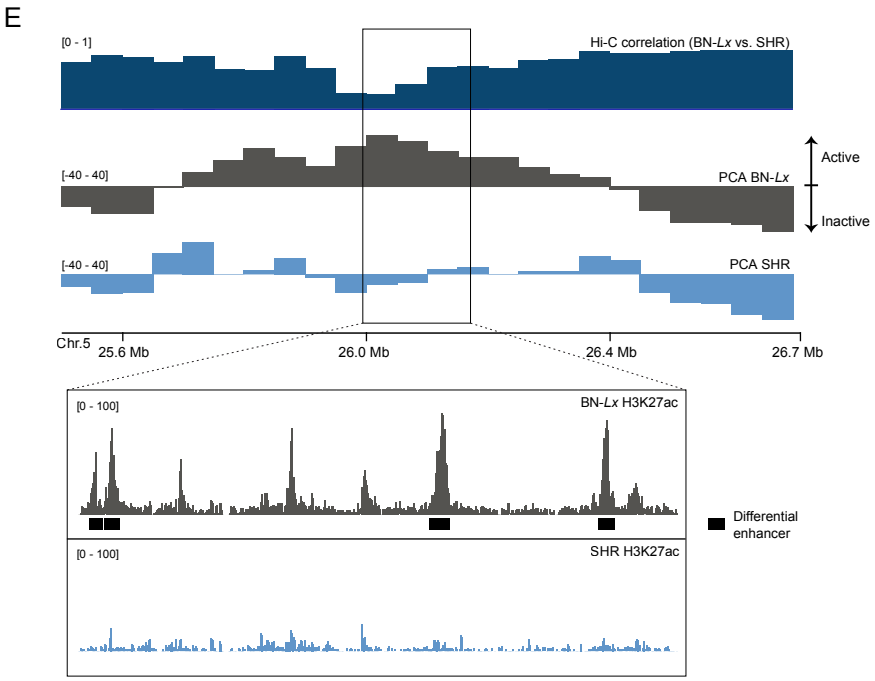
Requiring that at least one of the 50-kb windows in a region to be significantly different (Student's t-test, $p < 0.05$) resulted in a final set of 221 differential regions. Of these regions, 116 were unique to SHR, and 105 were unique to BN-*Lx* (Supplemental Table 5). Most of the differential domain regions were small, with a median size of 100 kb, and 78 of the regions comprised only one 50-kb window (Fig. 4F). In many cases, the differential domains were located at the border of the active domains, and an extension was observed in one of the strains (Fig. 4G). The largest region that was differential between the two strains was 1,25 Mb and positioned on chromosome 1. This region was only active in SHR and contained ten protein-coding genes (Fig. 4F and 4G). Overall, the results show that the 3D chromatin organization between the two genetic backgrounds was highly comparable, although changes were observed in specific regions.

3D chromatin changes correspond with epigenetically differential regions

Next, we investigated the relationships among gene expression variations, enhancers, and the differential higher-order chromatin domains (Fig. 5A). The comparison of the H3K27ac ratio in enhancers within the 221 differential domains shows that the level of enhancer marks corresponded with the domain structure. Enhancers in SHR-specific active domains had higher H3K27ac ratios in SHR than in BN-*Lx*, and vice versa (Student's t-test, $p < 4.9*10e-18$) (Fig. 5B). Surprisingly, a strong correlation was not observed for either the expression of genes with a TSS in a differential domain ($p < 0.14$) (Fig. 5B) or the expression levels of the closest TSS ($p < 0.13$). Thus, differential active domains corresponded to differences in enhancer levels, rather than to gene expression. To determine whether differential enhancers can alter 3D chromatin organization, we analyzed the Hi-C profiles around the 208 differential enhancers between BN-*Lx* and SHR (FDR < 0.05; log10(average CPM) > 2).

Figure 5. Differential chromatin domain organization correlates with differential enhancers in BN-*Lx* and SHR. (A) Schematic showing the position of the higher-order chromatin structure and epigenetics in the interaction network. (B) Ratios of H3K27ac levels in enhancers of differential higher-order domains. (C) Bar plot with correlation scores between BN-*Lx* and SHR for the 50-kb bins that contain differential or all enhancers (left) and differential or all genes (right). (D) 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. (E) Example showing the co-clustering of differential enhancers in a chromatin domain that was differentially active between BN-*Lx* and SHR. The zoomed region shows four enhancers that are referred to as "differential" between the two strains. Enhancers that show quantitative differences but are not significantly differential between the two strains can be seen.


A
Coding
Exon
Damaged protein
RNA degradation
Genomic variation
Noncoding
Enhancer
Gene expression change
Promoter
3D Chromatin organization
B
p < 4.9*10e-18
In BN-Lx-specific domains
In SHR-specific domains
p < 0.14
variance (log2(SHR / BNLx))
Enhancers
Gene expression
C
All enhancers
Differential enhancers
All expressed genes
Diff. expressed genes
Fraction of enhancers or genes
Hi-C correlation (BN-Lx vs. SHR)
D
All enhancers
Differential enhancers
Fraction of enhancers
0 - 25 kb
25 kb - 50 kb
50 kb - 75 kb
75 kb - 100 kb
100 kb - 150 kb
200 kb - 300 kb
300 kb - 400 kb
400 kb - 500 kb
500 kb - 600 kb
600 kb - 700 kb
> 700 kb
Distance to nearest active TSS

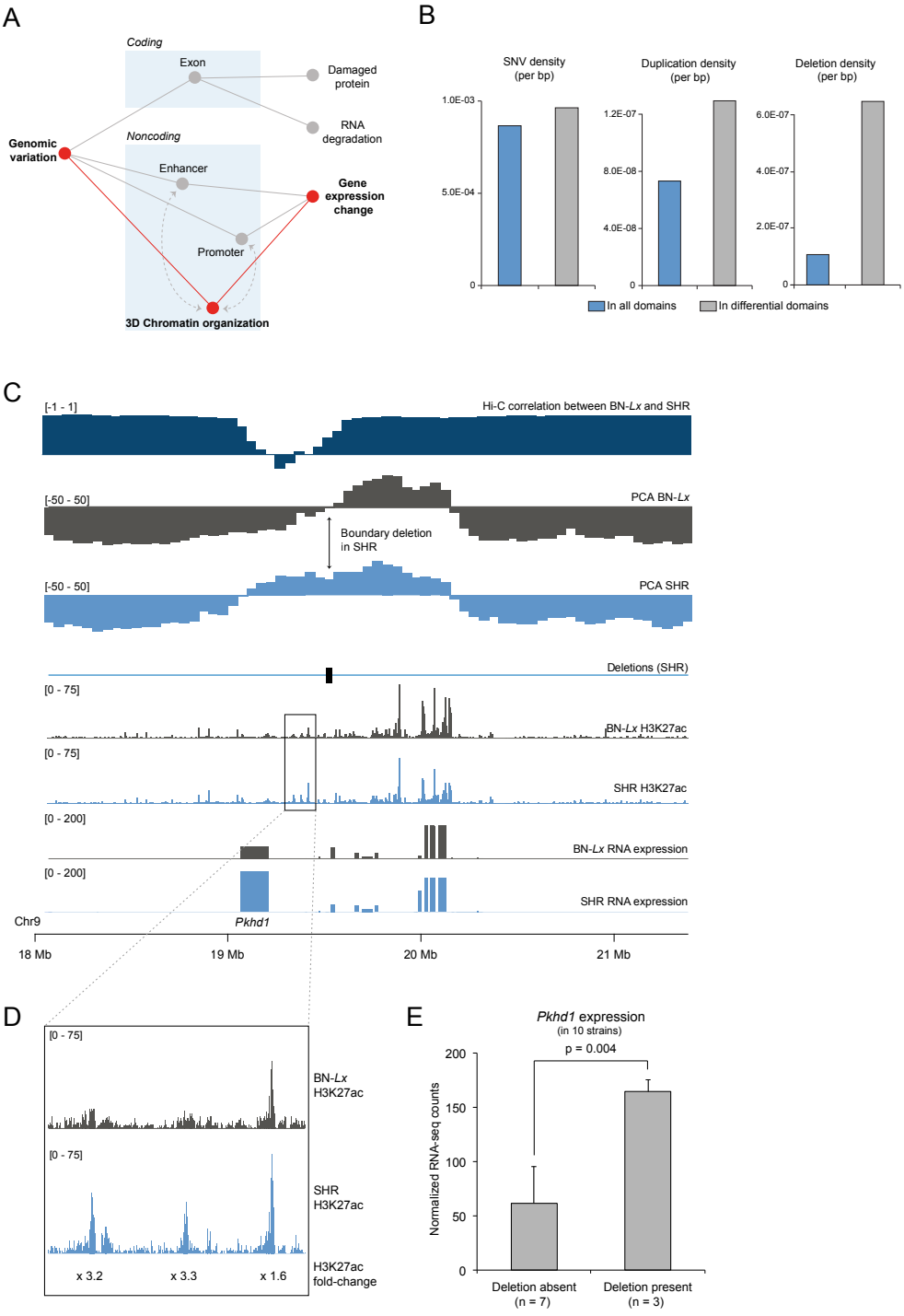

E
[0 - 1]
Hi-C correlation (BN-Lx vs. SHR)
[-40 - 40]
PCA BN-Lx
Active
Inactive
[-40 - 40]
PCA SHR
Chr.5
25.6 Mb
26.0 Mb
26.4 Mb
26.7 Mb
[0 - 100]
BN-Lx H3K27ac
Differential enhancer
[0 - 100]
SHR H3K27ac

The correlation scores between the Hi-C data of BN-*Lx* and SHR were lower in the 50-kb regions around the differential enhancers than in the set of all 50-kb bins that contained enhancers (Fig. 5C). A correlation > 0.9 was observed in 88% of the complete set of enhancers and 71% of the differential enhancers (for examples, see Supplemental Fig. 3). Analyzing the most differentially expressed genes also showed a drop in correlation values, but this was less pronounced than that for enhancers (Fig. 5C). The percentage of windows with a correlation >0.9 decreased from 89% to 81% around the TSS of differentially expressed genes. Interestingly, almost half of all 208 differential enhancers resided in inactive chromatin domains, whereas only one-fifth of the complete set of enhancers resided in these domains. This suggests that differential enhancers are more often located in inactive domains. In line with this, differential enhancers were often located relatively far away from the nearest active gene promoter (Fig. 5D). Fifty-six percent of the differential enhancers and 23% of the complete set of enhancers were positioned further than 100 kb from an active TSS. Differential enhancers that did localize to active domains frequently clustered within domains (Fig. 5E). Out of the 111 differential enhancers in the active domains, 42 (38%) mapped only to 18 out of the 941 active BN-*Lx* domains (2%). Our results show that a strong correlation between chromatin domain organization and the distribution of regulatory elements was observed. Domains with low correlation between strains possessed many differential enhancers that frequently co-localized with each other. In addition, the domain organization appeared to be linked more strongly to differences in enhancers, rather than to gene expression variation.

A genomic basis for differential chromatin organization

To assess the contribution of genomic variation to differences in 3D chromatin organization (Fig. 6A), we first determined the distribution of genomic variants from BN-*Lx* and SHR over the identified chromatin domains. For both SNVs and CNVs, a difference in the distribution over active and inactive domains was observed.

Figure 6. Differential Hi-C domains correlate with genetic variation. (A) Schematic showing the position of the higher-order chromatin structure and genetics in the interaction network. (B) Densities of genetic variants in all active domains versus differential domains. (C) Example of a differential domain containing a deletion. (D) Zoomed display of the three elevated enhancers in SHR (blue) compared to BN-Lx. Fold change was determined based on normalized H3K27ac read counts per enhancer. (E) Bar plot showing the gene expression levels of Pkhd1 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.


A
Coding
Exon
Damaged protein
RNA degradation
Noncoding
Genomic variation
Enhancer
Gene expression change
Promoter
3D Chromatin organization
B
SNV density (per bp)
Duplication density (per bp)
Deletion density (per bp)
1.0E-03
5.0E-04
0
1.2E-07
8.0E-08
4.0E-08
6.0E-07
4.0E-07
2.0E-07
In all domains
In differential domains
C
[-1 - 1]
Hi-C correlation between BN-Lx and SHR
[-50 - 50]
PCA BN-Lx
Boundary deletion in SHR
[-50 - 50]
PCA SHR
Deletions (SHR)
[0 - 75]
BN-Lx H3K27ac
[0 - 75]
SHR H3K27ac
[0 - 200]
BN-Lx RNA expression
[0 - 200]
SHR RNA expression
Chr9
Pkhd1
18 Mb
19 Mb
20 Mb
21 Mb
D
[0 - 75]
BN-Lx H3K27ac
[0 - 75]
SHR H3K27ac
x 3.2
x 3.3
x 1.6
H3K27ac fold-change
E
Pkhd1 expression
(in 10 strains)
p = 0.004
Normalized RNA-seq counts
200
150
100
50
0
Deletion absent (n = 7)
Deletion present (n = 3)

In the active domains, SNVs showed densities of 0.88 SNVs/kb versus 1.20 SNVs/kb in inactive domains (1.36-fold enrichment of SNVs in inactive chromatin domains). The CNV distribution was even more shifted towards inactive domains with only 27% mapping to active domains (128/342 of the deletions and 109/408 of the duplications). Therefore, genetic variation was lower in the active domains, which was in line with negative selection mechanisms for harmful variation in these regions. However, genetic variation was enriched in active domain regions that were differential between strains. The densities of SNV, duplication, and deletion exhibited 1.10-fold, 1.77-fold, and 6.2-fold increases, respectively (Fig. 6B and examples in Supplemental Fig. 4). In one example, the range of an active domain in SHR was extended by 450 kb, which was most likely a result of a deletion at the domain boundary that was present in BN-Lx (Fig. 6C). This extended active domain in SHR could serve as the basis of the observed increase in H3K27ac levels in three enhancers that were present in the differential locus (Fig. 6D). Also, the expression of the only expressed gene in the differential part of this domain, polycystic kidney and hepatic disease 1 gene (*Pkhd1*), was higher. To define if this particular deletion, which was located 155 kb away from the gene, could be causal for the increased expression of *Pkhd1* in the SHR strain, we determined the strain distribution pattern of the deletion and the expression of *Pkhd1* in all ten rats that were initially examined in this study. Every strain that carried the deletion allele also displayed increased *Pkhd1* expression (Fig. 6E). Although the deletion was likely to be causal for the increased expression of *Pkhd1*, we cannot exclude the involvement of other variants in that same genomic region.

Discussion

We have systematically studied variations at the genetic, epigenetic, and 3D structural levels and used integrated analyses to reveal how different regulatory levels inter-relate and affect gene expression.

Effects of genetic variation on the epigenetics of regulatory elements

We demonstrate that enhancers and promoters showed variability between genetic backgrounds and that part of the epigenetic variation could be

explained by underlying genetic variants. This is in line with recent studies showing histone marks as heritable traits [18, 19]. We show that part of the variation was caused by CNVs, which demonstrates the importance of surveying structural variation in genetic studies. The majority of the differential enhancers could not be explained by genetic variation and may result from differences in trans-regulatory mechanisms, rather than genetic variants inside the regulatory elements. Moreover, increasing the statistical power including more strains or performing the analyses at multiple time points during development may reveal stronger associations [27]. Nonetheless, the genetic contribution to epigenetic variation was evident. The SNVs that were linked to differences in H3K27ac levels showed enrichment in specific TF-binding motifs, suggesting that TF binding acts upstream of H3K27 acetylation. Several studies have examined the intricate relationship between the presence of SNVs and the level of TF binding. It has become evident that a SNV that disrupts a TF motif can result in the loss of TF binding, especially if the SNV is located in a very important position within the motif [11]. However, there are also many SNVs that are located in binding motifs but do not cause the loss of TF binding. This may be due to the presence of additional binding sites in close proximity [17, 19]. Furthermore, the loss of TF binding can occur independently of disruptions within the TF-binding motif, because DNA binding can depend on other factors [11, 28]. Although the multifactorial nature of TF binding to enhancers makes it difficult to use genomic data to predict the effects of individual SNVs, we have clearly illustrated a relationship among genetic variants, TF binding, and H3K27 acetylation in this study.

3D chromatin organization in different genetic backgrounds

We have previously analyzed the 3D chromatin organization in two rat strains that have a degree of genetic variation similar to the variation that is found between two unrelated human individuals [25]. Although the overall conservation of 3D structure was very high at the 50-kb resolution, some differences were detected between strains. It is possible that differences between smaller, specific interactions may be missed at this resolution. The high conservation of 3D structure is in agreement with a recent study that reported a high correlation in Hi-C data between different tissues and cell types and even across species [16]. However, there are few known examples of loci with differential 3D organization between cell types. These loci contain genes that reposition from the

inactive to the active compartment of the nucleus upon their upregulation during differentiation. Examples include *beta-globin*, which is expressed in erythroid cells [26], and *Ebf1*, which is upregulated during B-cell differentiation [29]. In addition, the introduction of a very strong enhancer can result in subtle changes in nuclear organization, such as the ectopic integration of the *beta-globin* Locus Control Region. This results in altered nuclear interactions that surround the integration site [30] but only within the limits set by the rigid gross chromosomal organization [31]. In this study, we show that naturally occurring variations in enhancers also corresponded with changes in nuclear interactions. However, differential gene expression between genetic backgrounds corresponded to a more limited change in 3D organization. These results indicate that the 3D organization around expressed genes is much more robust than around active enhancers. The regions in the genome that have marked differences in 3D organization between the genetic backgrounds had a high density of genetic variation, which suggests that a genetic component contributes to these differences. The genomic deletions and duplications are of particular interest. The targeted deletion of a domain boundary in the *Xist* locus was previously demonstrated to alter domain organization [32]. Here, we show that naturally occurring genetic variations may result in similar changes. Furthermore, this mechanism may result in heritable phenotypic variations.

Epigenetic and 3D dynamics of enhancers in the light of evolution

Enhancers contain a larger amount of genetic variants than promoters and also show more epigenetic variability between strains. The relatively low selective pressure on enhancers may be partially explained by the notion that enhancers often act in a combinatorial and partially redundant fashion. In line with this, the functional effects of variants that are located in enhancers often depend on the presence of additional variants in other enhancers in the same locus [33]. Our results show that differential enhancers were often located farther away from genes than non-differential enhancers. It is possible that enhancers in gene-poor regions cannot affect gene expression regulation, thus allowing epigenetic variation to occur at these sites. Importantly, dynamics in enhancers have led to large and crucial evolutionary changes. For example, a single homeobox (*Hox*) polymorphism in snakes results in the functional loss of a *Hox* enhancer and a consequent change in rib formation [34]. Moreover,

changes in the spatial interactions of enhancers have resulted in new evolutionary developments. The regulation of *Hoxd*, which forms the basis of digit formation, arose during evolution through the establishment of novel interactions between enhancers in two neighboring topological domains [35]. Although our data show that a certain level of variability in enhancers and concomitant 3D organization appeared to be allowed, these types of variation can ultimately result in functionally important changes in gene expression and evolution.

Overall, results from our systematic analyses suggest transcriptional robustness and variability in underlying regulatory networks in different genetic backgrounds. This integrative approach has demonstrated the complexity of the effects of single nucleotide and structural genomic variations. However, further multilevel experimental analyses are required for the proper understanding of the functional effects of individual genetic variants.

Methods

Liver tissue collection

We obtained snap-frozen liver tissues from ten inbred rat strains (6 weeks old). Liver tissues from six out of ten strains were kindly provided to us by Dr. James D. Shull (ACI/SegHsd; University of Wisconsin, Madison, WI, USA), Dr. Myrna Mandel (M520/N, MR/N, and WN/N; NIH - Office of Research Services, Bethesda, MD, USA), and Dr. Michal Pravenec (BN-Lx/Cub and SHR/Olalpcv; Charles University, Prague). Tissues from F344/NCrHsd, WKY/NHsd, BN/SsNOlaHsd, and BUF/SimRijHsd strains were purchased from Harlan Laboratories (Horst, The Netherlands).

Preparation of cross-linked cell nuclei for ChIP and Hi-C

Snap-frozen and powdered rat liver tissues (~40 mg) were resuspended in 2 mL cold phosphate-buffered saline (PBS)-10% fetal calf serum (FCS) and dissociated using a 40 μm nylon cell strainer (BD Biosciences, New Jersey USA). The cell suspension was cross-linked with 2% formaldehyde in a total volume of 10 mL PBS/10%FCS for 10 minutes at RT with rotation (20°C). Glycine (0.125 M) was added to quench the reaction, and the cells were stored on ice. Following the cross-linking procedure,

samples were centrifuged for 8 minutes at 400 g at 4°C. Pelleted cells were washed with 1 mL cold PBS and centrifuged again at 400 g, 4°C, for 5 minutes. After removing the supernatant, the cell pellet was dissolved in 1 mL freshly prepared lysis buffer (50 mM Tris pH 7.5, 150 mM NaCl, 5 mM ethylenediaminetetraacetic acid [EDTA], 0.5% nonyl phenoxylpolyethoxylethanol-40, 1% Triton X-100, and 1× complete, EDTA-free Protease Inhibitor Cocktail [#11873580001, Roche Applied Sciences, Indianapolis, IN, USA]) to recover cross-linked nuclei. Cells were lysed for 10 minutes on ice, and methyl green-pyronin staining was used to determine complete cell lysis. After lysis completion, nuclei were washed twice in cold PBS.

ChIP, library preparation, and sequencing

Cross-linked nuclei were dissolved in 100µL lysis buffer (MAGnify system, Invitrogen, Carlsbad CA, USA) with 1x protease inhibitors (MAGnify system, InvitrogenTM) 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 was used at an amount equivalent to ±20 mg input liver tissue (150–300 bp) for immunoprecipitation (IP). H3K4me3 and H3K27ac IPs were carried out using the MAGnify system (InvitrogenTM, 49-2024), following the manufacturer's 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 performed, according to the manufacturer's instructions. Libraries were sequenced on SOLiD 5500XL Wildfire, thus resulting in 40-bp reads.

Calling enhancer and promoter regions

Sequencing reads were mapped using the Burrows-Wheeler Aligner (BWA-0.5.8c) (settings: –c –l 25 –k 2 –n 10) [36] to the reference genome, RGSC3.4. This resulted in 67-71 million mapped reads per sample for the H3K27ac ChIP and 14-21 million mapped reads per sample for the H3K4me3 ChIP. Peak calling was done using the Model-based Analysis of ChIP-Seq [37] (version 1.4, settings: -g 2718897334 -B -S --to-small -p 1e-10 bandwidth=300 model=TRUE shiftsize=100), and the input

DNA for chromatin was used as the control. Called peaks for H3K4me3 and H3K27ac were processed separately. Peak regions of all strains were merged using the mergeBed command from the BEDtools suite [38], 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, whereas those that did overlap an H3K4me3 peak were assigned as promoters. The strain specificity for each enhancer and promoter was determined by assessing whether each strain possessed an H3K27ac or H3K4me3 peak that overlapped 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 [38]. 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. Differential enhancers between BN-*Lx*/Cub and SHR/OlaIpcv were determined using the R-package EdgeR, and both strains were used to calculate the common dispersion. The threshold was set at FDR < 0.05. Because enhancers with a low number of sequencing reads are noisy, we also set a threshold to the minimum amount of reads required. Based on the distribution in the MA plot, as visualized with the plotSmear function in the EdgeR package, we required that log10 (average CPM) > 2.

SNV calling

To make a comprehensive comparison between rat strains, we applied multi-sample SNV calling using the Genome Analysis Toolkit (GATK) [39]. SNVs were determined using raw data, as described previously [24, 25]. Nine strains (ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WN/N, WKY/N, and BN-*Lx*/Cub) were sequenced on the SOLiD™ platform, and they were analyzed simultaneously, using multi-sample SNV calling by the haplotype caller in GATK. The LE/Stm rat strain was included to determine specificity, and SNP array data were used to determine sensitivity, as described previously [24]. Based on 2.1 Mb of bacterial artificial chromosome sequences that were available for the LE/Stm strain, SNV calls were 99% specific. Based on SNP array data, SNVs were 99.5% sensitive. The SHR/OlaIpcv strain could not be included in the multi-sample calling, because it was 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 was slightly less than the other nine strains, because we were unable to take advantage of multi-sample calling to increase sensitivity. SNVs for all ten strains were then merged into one variant call format file.

Association of H3K27ac and gene expression 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 strain. 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 separated the ten strains in two groups of at least three strains, such that a Student's t-test could be performed. The final set that was 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 for two-sided testing. For an accurate p-value per SNV, multiple testing corrections would be required. However, the p-values were merely used as the threshold above which we could demonstrate enrichment in the real versus the randomized set. Intersections between CNVs and enhancers were determined using the intersectBed command form in the BEDtools suite [38].

Motif analyses

TF-binding motifs were called using Find Individual Motif Occurrences version 4.9.1 and JASPAR CORE 2014 vertebrates with the p-value threshold of 1e-4. Motifs were called on the SNVs that were located within an enhancer or promoter and used in the association analyses (see above). The sequence that was used for each SNV contained the 10 bp upstream to the SNV, the SNV position itself, and the 10 bp downstream of the SNV (21 bp in total). Motifs that did not cover the SNV were filtered out in a secondary step. These included motifs that were shorter than 11 bp, ended before position 11, or ended after position 11. The reference and alternative alleles were both tested for overlap with TF motifs and then combined to determine whether the SNV was located in a TF motif. If both alleles were located within a motif, it still counted as one occurrence. Chi-square tests were used to test the enrichments of specific TF motifs in the SNVs that were associated with H3K27ac levels. The frequency of the occurrence of TF motifs in the total set of SNVs tested was set

as the expectancy value in a Chi-square test. The tested frequency was the occurrence of the TF motif in the set of SNVs that associated with H3K27ac levels with $p < 0.05$ (see above and Fig. 2). For a more exact p-value, multiple testing would be required. However, our goal was to find the differences between different TFs and to demonstrate that the enrichment of liver TFs was stronger than other TFs.

Hi-C: massive parallel sequencing of proximity-based ligation products

Isolated and cross-linked liver nuclei of three BN-*Lx*/Cub and three SHR/OlaIpcv rats were digested with the DpnII restriction enzyme (#R0543, NEB, Ipswich, MA, USA). Subsequently, the proximity ligation of interacting fragments was performed using T4 DNA ligase (#10799009001, Roche Applied Sciences) to produce the 3C template, according to a previously described protocol [26]. After reverse cross-linking and precipitation, 10 µg of the template was sheared in microtubes (AFA Fiber Pre-Slit Snap-Cap 6×16mm, 520045) using the Covaris S2 sonicator (1 cycle of 25 seconds; duty cycle: 5%, intensity: 3, cycles per burst: 200, frequency sweeping). Fragments that ranged in size from 500 to 1500 bp were selected using a 2% agarose gel. Size-selected fragments (1.1 µg) were used as the 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 and 950 bp. These libraries were sequenced in a paired-end manner on the Illumina HiSeq 2500, resulting in 2×100-bp reads. Sequenced read pairs were mapped using Burrows-Wheeler Aligner (BWA-0.7.5a) (settings: bwa mem -c 100 -M) [36] to reference genome RGSC3.4, thus yielding 70 million 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 [40]. Sequenced read pairs were filtered to have minimum distance of 1.5 kb to omit self-ligated fragments. Full filter settings using Homer makeTagdirectory were: -update -removePEbg -fragLength 1500 -removeSpikes 10000 5. This resulted in 60 million usable read pairs per strain (three replicates combined). PCA analyses and correlation differences between strains were calculated using window sizes of 100 kb, with a step size of 50 kb (a 100-kb 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

Snap-frozen and powdered liver tissues (~40 mg) were 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)). Differentially expressed genes between BN-*Lx*/Cub and SHR/OlaIpcv strains were determined using the R-package EdgeR. Read counts that were normalized to EdgeR used both strains to calculate the common dispersion. The threshold was set at FDR < 0.05. Because genes with a low number of sequencing reads generate a lot of noise, we also set a threshold to the minimum amount of reads required. Based on the distribution in the MA plot, as visualized with the plotSmear function in the EdgeR package, we required that log10(average CPM) > 2.

Data access

The ChIP-seq, RNA-seq, and Hi-C data are available at the European Nucleotide Archive/Sequence Read Archive under accession number PRJEB6392.

Supplemental information

All supplemental files in this chapter can be downloaded from http://www.hubrecht.eu/research/cuppen/hermsen_thesis.html

Acknowledgments

This work was financially supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under the No. HEALTH-F4-2010-241504 (EURATRANS) and NWO-CW TOP (700.58.303) grants to EC. MS acknowledges funding from the NWO Vernieuwingsimpuls program (grant number 863.10.007). PF was supported by EURATRANS and EMBL. We are grateful to Dr. James D. Shull (University of Wisconsin, Madison), Dr. Myrna Mandel (National Institute of Health [NIH] - Office of Research Services), and Dr. Michal Pravenec (Charles University, Prague) for kindly providing liver tissues. We thank Geert Geeven and Peter Krijger for their helpful suggestions about Hi-C experiments.

Author 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. PT, EdB, and MV performed next-generation sequencing. WS and SB performed next-generation sequence mapping and SNP calling. WdL contributed to Hi-C library construction and data analysis. DT and PF contributed to scientific discussions regarding ChIP-seq experiments and TF overlaps. SvH, RH, EC, and MS conceptually designed the study, coordinated the experiments, critically discussed the results, and wrote the manuscript. All authors have read and approved the final version of the manuscript.

References

1. 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. Proceedings of the National Academy of Sciences of the United States of America 2009, 106(23):9362-9367.
2. Kumar V, Wijmenga C, Withoff S: From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Seminars in immunopathology 2012, 34(4):567-580.
3. Pennisi E: The Biology of Genomes. Disease risk links to gene regulation. Science 2011, 332(6033):1031.
4. 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(6822):860-921.
5. 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(5507):1304-1351.
6. 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(7414):57-74.
7. 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(7249):927-930.
8. Ponting CP, Hardison RC: What fraction of the human genome is functional? Genome research 2011, 21(11):1769-1776.

9. Ward LD, Kellis M: Interpreting noncoding genetic variation in complex traits and human disease. Nature biotechnology 2012, 30(11):1095-1106.

10. Lehner B: Genotype to phenotype: lessons from model organisms for human genetics. Nature reviews Genetics 2013, 14(3):168-178.

11. 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(3):530-540.

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. Proceedings of the National Academy of Sciences of the United States of America 2010, 107(50):21931-21936.

13. 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. Nature genetics 2007, 39(3):311-318.

14. 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(5950):289-293.

15. 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. Nature genetics 2011, 43(7):630-638.

16. 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(7398):376-380.

17. 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(7477):487-492.

18. 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(6159):750-752.

19. 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(6159):744-747.

20. 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(6159):747-749.

21. 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. Nature methods 2010, 7(4):248-249.

22. Ng PC, Henikoff S: SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research 2003, 31(13):3812-3814.

23. Weischenfeldt J, Symmons O, Spitz F, Korbel JO: Phenotypic impact of genomic structural variation: insights from and for human disease. Nature reviews Genetics 2013, 14(2):125-138.

24. 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. Nature genetics 2013, 45(7):767-775.

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 biology 2012, 13(4):r31.

26. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W: Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nature genetics 2006, 38(11):1348-1354.

27. Francesconi M, Lehner B: The effects of genetic variation on gene expression dynamics during development. Nature 2014, 505(7482):208-211.

28. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE: Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature 2009, 462(7269):65-70.

29. Lin YC, Benner C, Mansson R, Heinz S, Miyazaki K, Miyazaki M, Chandra V, Bossen C, Glass CK, Murre C: Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. Nature immunology 2012, 13(12):1196-1204.

30. Noordermeer D, Branco MR, Splinter E, Klous P, van Ijcken W, Swagemakers S, Koutsourakis M, van der Spek P, Pombo A, de Laat W: Transcription and chromatin organization of a housekeeping gene cluster containing an integrated beta-globin locus control region. PLoS genetics 2008, 4(3):e1000016.

31. Noordermeer D, de Wit E, Klous P, van de Werken H, Simonis M, Lopez-Jones M, Eussen B, de Klein A, Singer RH, de Laat W: Variegated gene expression caused by cell-specific long-range DNA interactions. Nature cell biology 2011, 13(8):944-951.

32. 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(7398):381-385.

33. 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 research 2014, 24(1):1-13.

34. Guerreiro I, Nunes A, Woltering JM, Casaca A, Novoa A, Vinagre T, Hunter ME, Duboule D, Mallo M: Role of a polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine. Proceedings of the National Academy of Sciences of the United States of America 2013, 110(26):10682-10686.

35. 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(6137):1234167.

36. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25(14):1754-1760.

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

38. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010, 26(6):841-842.

39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 2010, 20(9):1297-1303.

40. 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(4):576-589.

5

Chapter 5

Lack of major genome instability in tumors of p53 null rats

Roel Hermsen[1], Pim Toonen[1], Sameh A. Youssef[2], Raoul Kuiper[2], Ewart Kuijk[1], Alain de Bruin[2], Edwin Cuppen[1] and Marieke Simonis[1]

1 Hubrecht Institute, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands.

2 Dutch Molecular Pathology Center. Department of Pathobiology, Faculty of Veterinary Medicine, Utrecht University, Yalelaan 1, 3584 CL Utrecht, The Netherlands.

Abstract

Tumorigenesis is often associated with loss of tumor suppressor genes (such as *TP53*), genomic instability and telomere lengthening. Previously, we generated and characterized a rat p53 knockout model in which the homozygous rats predominantly develop hemangiosarcomas whereas the heterozygous rats mainly develop osteosarcomas. Here we use this p53 mutant rat model to investigate the integrity of the tumor genomes. We find that the tumors that arise in the heterozygous and homozygous *Tp53* genotype are different in both tumor type and genomic stability. We investigated the p53 status and of tumors in heterozygous and homozygous $Tp53^{C273X}$ knockout rats and found that in all tumors p53 is fully inactivated. Strikingly, when we investigated genomic integrity, tumors from homozygous animals show very limited aneuploidy and a very low amount of copy-number variants affected base pairs compared to the tumors from heterozygous animals. Also complex structural rearrangements such as chromothripsis and breakage-fusion-bridge cycles were never found in tumors from homozygous animals, while these where readily detectable in tumors from heterozygous animals. Finally, we measured telomere length and telomere lengthening pathway activity and found that tumors of homozygous animals have longer telomeres but do not show clear telomerase or alternative lengthening of telomeres (ALT) activity differences as compared to the tumors from heterozygous animals. Taken together our results demonstrate that host p53 status has a large effect on genomic instability characteristics, where full loss of functional p53 is not the main driver of large-scale structural variations. Our results also suggest that chromothripsis primarily occurs under p53 heterozygous rather than p53 null conditions.

Introduction

p53 is regarded as an important gate keeper of genome integrity in healthy cells. In the majority of all tumors the *TP53* gene is mutated and the p53 protein is inactivated [1, 2]. Inherited mutations in *TP53* are the cause of Li-Fraumeni syndrome, which is marked by a high incidence of tumors and the development of cancer early in life [3]. Genomic instability can result in different types of large structural rearrangements in the genome.

A recent analysis of the Sonic-Hedgehog medulloblastoma brain tumors of Li-Fraumeni patients indicated a strong, permissive connection between *TP53* mutations and chromothripsis, i.e. a massive complex chromosomal rearrangement that occurs in a single event [4]. The exact mechanism of chromothripsis remains elusive but evidence suggests it may occur during cell division [5]. In cancer, cell cycling is often deregulated and accelerated and one of the effects of this rapid cell cycling is telomere shortening. When telomeres become critically short, chromosome end-to-end fusions can occur recurrently, which may result in breakage-fusion-bridges (BFBs) [6]. To maintain telomere length, telomerase or the alternative lengthening of telomeres (ALT) pathway is often activated in tumors and in other cell types that undergo a high rate of cell cycling, like stem cells [7]. Previously, we generated and characterized a rat p53 knockout model in which the homozygous mutant animals completely lack the p53 protein as a result of a C273X nonsense mutation in the sixth exon, which truncates the protein at the DNA binding domain, hereby eliminating functionally essential domains including the nuclear localization domain and the homo-oligomerization domain. The complete absence of functional p53 protein in homozygous mutant animals was demonstrated in two ways: First upon ultraviolet radiation no full-length or truncated p53 could be detected in isolated homozygous mutant REFs. Secondly a Western blot analysis on REF lysates treated with doxorubicin confirms complete lack of full-length or truncated p53 in homozygous mutant cells [8]. In this model, the homozygous rats predominantly develop hemangiosarcomas whereas the heterozygous rats mainly develop osteosarcomas [8]. This shows that tumor development depends on genotype, but how genotype affects the stability of tumor genomes remains unclear.

In this study we investigate the integrity of the tumor genomes of the p53 knockout rat model, comparing heterozygous and homozygous genotypes. Strikingly, we find that tumors that are formed in the homozygous p53 knockout animals have a stable tumor genome and are fundamentally different from tumors formed in heterozygous animals, which display extensive chromosomal aberrations and complex structural rearrangements.

Material and Methods

Animals

All experiments were approved by the Animal Care Committee of the Royal Dutch Academy of Sciences according to the Dutch legal ethical guidelines. Experiments were designed to minimize the number of required animals and their suffering. The *Tp53* knockout rat (Crl:WI(UL)-Tp53$^{m1/Hubr}$) was generated by target-selected ENU-driven mutagenesis (for detailed description, see [9]). The heterozygous-mutant animal was outcrossed several times to eliminate confounding effects from background mutations induced by ENU. *Tp53*^{C273X} rats were bred under standard conditions to generate animals that are heterozygous and homozygous for this allele. Animals were housed under standard conditions in groups of two to three per cage per sex under controlled experimental conditions (12-hour light/dark cycle, 21 ±1°C, 60% relative humidity, food and water ad libitum). Health status and tumor development was monitored weekly. After detection of a visible tumor (age 8-71 weeks) animals were sacrificed.

Tissue sampling

The tumor was dissected and split into two: one sample was fixed in 4% neutral buffered formalin for histological analysis and another was snap frozen sample in liquid nitrogen for molecular analysis. DNA extraction was carried out with 20 to 30 mg of snap-frozen and ground tissue using Genomic tips (Qiagen, Venlo, Netherlands). DNA quality was assessed on an agarose gel. If the DNA was degraded, the isolation was repeated. If samples were contaminated with protein or RNA, they were re-isolated using DNeasy spin columns (Qiagen) to ensure high purity of the DNA essential for aCGH and NGS.

Histological analysis

Immunohistochemistry (IHC) was performed only on those tumors that were undifferentiated on hematoxylin and eosin (HE) stain. The IHC staining procedures were performed according to the protocol of the primary antibody manufacturer. SMA (Abcam (Cambridge, UK) at 1:100), Vimentin (Santa Cruz Biotech (Dallas, TX, USA) at 1:30), Factor VIII (Dako (Santa Clara, CA, USA), 1:500), and Osteocalcin (Merck Millipore (Billerica, MA, USA), 1:400), MyoD-1 (Dako) antibodies were applied overnight in humid chamber at room temperature, then sections were

washed and then incubated in (3,3'-diaminobenzidine (DAB) substrate kit for 4 minutes (Vector SK-4100, Vector Laboratories (Burlingame, CA, USA)). Positive and negative control sections were performed for every antibody.

Array Comparative Genomic Hybridization

Array CGH experiments were performed as previously described [10]. Data is available at [NCBI-GEO:GSE55895]. The oligonucleotide design, array fabrication, DNA labeling, hybridization were performed according to manufacturer's instructions. Microarrays were scanned with a G2565CA scanner (Agilent, Santa Clara, CA, USA) at resolution 2 μm, double pass. Microarray slides were reused once by removing the hybridized probe from the slide (after image acquisition) using the NimbleGen Array Reuse Kit (Roche Nimblegen, Basel, Switzerland). Image analysis was performed using Feature extraction software version 10.5.1.1 (Roche Nimblegen) and CGH-segMNT in NimbleScan 2.6 software (Roche NimbleGen).

Gains and losses

Cutoff values for log2 ratio were determined using a control versus control array. Sex and residual chromosomes were excluded from analysis. We defined chromosomal arm-level alteration as a single alteration or an aggregate of alterations that encompass >75% of a chromosomal arm.

Chromothripsis

We inferred chromothripsis when we observed at least 10 changes in segmental copy-number involving two or three distinct copy-number states on a single chromosome [4].

Breakage-fusion-bridges

We inferred BFB cycles when we observed a stair-like increase in copy-number with multiple segments (minimum of 3 increasing copy-number states) on one side and a single sharp drop on the other side; located at the chromosome start or end.

Telomere length

500ng of genomic DNA of each tumor was used for SOLiD 5500xl Wildfire (Thermo Fisher Scientific, Life Technologies, Waltham, MA, USA) library preparation and whole genome sequencing resulting in 0.10-1.34x coverage per sample. Reads were mapped onto a telomeric repeat

reference ($TTAGGG_{10}$) with BWA 0.5.9 [11] and average telomere length was calculated. Data is available at [NCBI-SRA:PRJEB5836].

Telomerase and ALT activity

Telomerase activity of tumor cell extracts measured by TRAPeze (TRAPeze® XL Telomerase Detection Kit, Merck Millipore). ALT activity of tumor cell extracts was measured previously described [12] by qPCR of CC assay product, normalized for the ALT-positive control cell line U-2 OS.

Results and Discussion

We investigated the genetic status of *Tp53* in all tumors found in heterozygous animals by capillary sequencing and found that all tumors show loss of heterozygosity (LOH); data not shown. Thus, all tumors in heterozygous animals have lost the functional *Tp53* allele and have become *Tp53* null.

Next, tumors were histologically examined to determine the tumor-type. In total we excised 26 tumors (10 $Tp53^{-/-}$, 16 $Tp53^{+/-}$) of which 25 could be classified as specific tumor types. In heterozygous animals we found 10 osteosarcomas, 2 fibrosarcomas, 1 leiomyosarcoma, 1 rhabdomyosarcoma and 1 transitional cell carcinoma. In homozygous animals we found 7 hemangiosarcomas, 2 fibrosarcomas and 1 leiomyosarcoma (Table 1). So even though the tumors of all animals do not have functional p53, the tumor spectrum is different between $Tp53^{+/-}$ and $Tp53^{-/-}$ animals. Our results corroborates our previous observation that $Tp53^{C273X}$ mutant rats mainly develop sarcomas and that the tumor spectrum of $Tp53^{+/-}$ and $Tp53^{-/-}$ animals is different. This shows that the timing of p53 loss affects tumor type.

To investigate whether absence of p53 results in high genomic instability in the tumors, we performed array comparative genomic hybridization (aCGH) on tumor versus control tissue to detect copy number variants (CNVs) (Supplementary figure 1). First we investigated the presence of aneuploidies in the tumors. We defined an aneuploidy as an event in which at least 75% of the chromosome or chromosomal arm was affected and the copy-number status throughout the affected region was the same.

Table 1. Overview of the characteristics of the tumors in mutant Tp53 rats in this study.

Rat ID	Genotype	Sex	Age (weeks)	Tumor diagnose	Immunohistochemistry	LOH	LOH type	Chromothripsis	Chromothripsis aff. chr.*	BFB cycles	BFB aff. chr.*	Amplified oncogene(s)
196	- / -	Male	21	fibrosarcoma	VM^{+}, SMA^{-}, OC^{-}, F/VIII^{-}	-	-	NO	-	NO	-	-
198	- / -	Male	13	fibrosarcoma	VM^{+}, SMA^{-}, OC^{-}, F/VIII^{-}	-	-	NO	-	NO	-	-
164	- / -	Male	13	hemangiosarcoma	VM^{+}, F/VIII^{+}	-	-	NO	-	NO	-	-
174	- / -	Male	18	hemangiosarcoma	ND	-	-	NO	-	NO	-	-
199	- / -	Male	14	hemangiosarcoma	ND	-	-	NO	-	NO	-	-
223	- / -	Female	15	hemangiosarcoma	ND	-	-	NO	-	NO	-	-
261	- / -	Male	12	hemangiosarcoma	ND	-	-	NO	-	NO	-	-
225	- / -	Male	10	hemangiosarcoma	F/VIII^{+}	-	-	NO	-	NO	-	-
247	- / -	Male	8	hemangiosarcoma	F/VIII^{+}	-	-	NO	-	NO	-	-
201	- / -	Male	20	leiomyosarcoma	VM^{+}, SMA^{+}, F/VIII^{-}	-	-	NO	-	NO	-	-
103	+ / -	Male	54	fibrosarcoma	ND	Yes	deletion	YES	6	NO	-	-
170	+ / -	Male	45	fibrosarcoma	VM^{+}, SMA^{-}, OC^{-}, F/VIII^{-}	Yes	deletion	YES	1	YES	7	*Myc*
153	+ / -	Male	49	leiomyosarcoma	VM^{+}, SMA^{+}, OC^{-}, F/VIII^{-}	Yes	affected by chromothripsis	YES	10**,13	YES	9	*Vegfa*
20	+ / -	Female	45	osteosarcoma	ND	Yes	gene conversion or copy-neutral SV	NO	-	NO	-	-
82	+ / -	Male	45	osteosarcoma	ND	Yes	gene conversion	YES	7	NO	-	-
97	+ / -	Male	47	osteosarcoma	ND	Yes	affected by chromothripsis	YES	3,8,10**	NO	-	-
100	+ / -	Female	45	osteosarcoma	ND	Yes	gene conversion or copy-neutral SV	NO	-	NO	-	-
112	+ / -	Female	71	osteosarcoma	ND	Yes	affected by chromothripsis	YES	3,10**	YES	6, 10	*Mycn, Alk*
118	+ / -	Male	41	osteosarcoma	ND	Yes	gene conversion	NO	-	NO	-	-
142	+ / -	Male	54	osteosarcoma	ND	Yes	deletion	YES	5	YES	9	*Vegfa*
144	+ / -	Male	68	osteosarcoma	ND	Yes	deletion	NO	-	YES	5, 7	*Myc*
160	+ / -	Female	52	osteosarcoma	ND	Yes	deletion	NO	-	NO	-	-
202	+ / -	Male	34	osteosarcoma	ND	Yes	deletion	YES	10,17	NO	14	-
9	+ / -	Female	36	rhabdomyosarcoma	MYD^{+}, F/VIII^{+}	Yes	gene conversion or copy-neutral SV	YES	6	NO	-	-
125	+ / -	Female	46	transitional cell carcinoma	ND	Yes	gene conversion or copy-neutral SV	NO	-	NO	-	-
95	+ / -	Male	54	ND	ND	Yes	deletion	NO	-	NO	-	-

* Affected chromosomes

** *Tp53* locus affected

ND: not determined, LOH: loss of heterozygosity, VM: Vimentin, OC: Osteocalcin, SMA: smooth muscle actin, F/VIII: factor VIII, MYD: MyoD-1. +: Positive IHC staining, —: negative IHC staining

We found a clear difference in the number of aneuploidy events between tumors that arise in heterozygous rats compared to those in homozygous rats: contrary to our expectations, tumors from homozygous animals show almost no deleted or gained chromosomes whereas tumors from heterozygous animals display massive aneuploidies (Figure 1a). In addition, we counted all base pairs that are affected by CNVs (excluding the aneuploidies) in tumors from homozygous animals to tumors of heterozygous animals. Strikingly, we find that in homozygous tumors, in addition to the absence of chromosomal events, also less base pairs are affected by CNVs (Figure 1b). In heterozygous rats, we observed highly aneuploid and CNV-rich genomes in osteosarcomas, in line with known human osteosarcoma genomic profiles [13]. Although the main tumor type observed in homozygous animals, hemangiosarcomas, is also reported to harbor many aneuploidies in human [14], we identified only three gained or lost chromosomes in seven hemangiosarcomas. A typical example of the type and amount of aberrations observed in osteosarcomas and hemangiosarcomas is shown in Figure 1c. These data demonstrate that genomic instability is not a prerequisite for tumor formation when p53 is already completely inactivated, but also that complete absence of p53 does not systematically induce large-scale genomic instability.

Recently, a link between *TP53* mutation status and the presence of chromothripsis has been postulated [4]. Therefore we set out to explore the presence of chromothripsis events in the tumors under study. In chromothripsis, one or a few genomic regions in the genome are randomly shattered and reassembled. To infer the occurrence of chromothripsis, we used the definition put forward by Rausch *et al* [4], which requires at least ten changes in segmental copy-number involving two or three distinct copy-number states on a single chromosome. In nine out of sixteen tumors from heterozygous rats we find evidence of chromothripsis (Figure 2a). In contrast, none of the tumors from homozygous rats display any evidence for chromothripsis. Our results support the previous finding that heterozygous p53 mutations increase the incidence of chromothripsis. Question remains what the order of events is; is loss of the functional *Tp53* allele necessary to elicit genomic instability, which then results in chromothripsis? The finding that none of the homozygous tumors display chromothripsis suggests that complete loss of p53 is not the main driver to elicit this type of genomic rearrangement. From our data, it appears more likely that the loss of one allele of p53 makes cells more sensitive

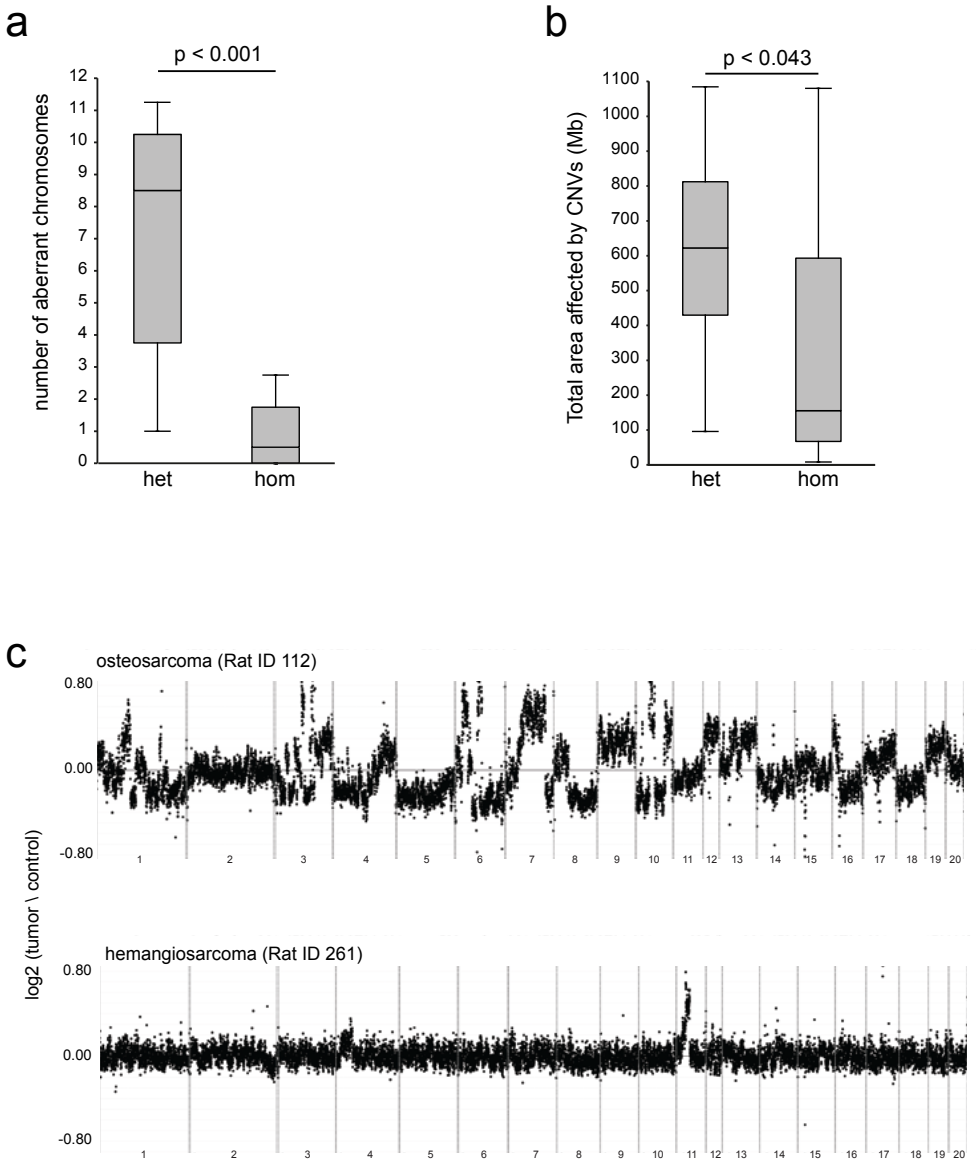


Figure 1. Tumors from homozygous *Tp53*^{C273X} mutant rats show significantly less CNVs than tumors from heterozygous *Tp53*^{C273X} rats. (a) Counts of fully gained or lost chromosomes (>75%) from array CGH data on tumors DNA versus healthy control DNA. Tumors from heterozygous animals (n=16) and tumors from homozygous animals (n=10) were compared using a Student's t-test. (b) A sum of all altered base pairs (gained or lost) per tumor. Tumors from heterozygous animals (n=16) and tumors from homozygous animals (n=10) were compared using a Student's t-test. (c) Two representative examples of the aCGH data. The upper panel shows the amount of CNVs in the genome of an osteosarcoma from a heterozygous animal (Rat ID 112); the lower panel shows CNVs in a hemangiosarcoma from a homozygous animal (Rat ID 261).

to (some forms) of genomic instability. Resulting rearrangements subsequently result in loss of the healthy p53 allele and further tumor development. In support of this view we find in the aCGH data that in three out of nine chromothripsis cases the second *Tp53* allele is lost via this mechanism (Table 1).

Next we studied another hallmark of cancer and genomic instability: breakage-fusion-bridges (BFB) cycles, including gene amplifications. As a consequence of BFB cycles a distinct stair-like gradient of copy-number increasing genomic segments are formed which can be detected by array CGH [15]. Exploration of our data for such signatures identified BFB cycle events in six out of sixteen tumors formed in heterozygous rats (Figure 3a). All chromosomes with a BFB cycle-signature contained highly amplified segments with well-known oncogenes like *Myc*, *Mycn*, *Vegfa* and *Alk* (Figure 3b and Table 1). In tumors from homozygous animals however, we did not find any evidence for BFB cycles. BFB cycles are caused by telomeric shortening and subsequent crisis. We hypothesized that the relative stability of p53 null tumors depends on activated telomere-lengthening mechanisms that protect against BFBs in the tumors of homozygous animals. Therefore, we interrogated telomere length of the tumor cells by next-generation sequencing. We counted and normalized the reads that mapped to the telomeric repeat (TTAGGG) and calculated the length of an average telomere per sample. Tumors of homozygous animals showed significantly longer telomeres compared with tumors of heterozygous animals ($p<0.05$) or with ear control DNA ($p<0.01$, Student's t-test) (Figure 4a). This suggests that telomere length may indeed be a factor that is involved in the difference in genomic rearrangements. This difference can be caused by increased shortening in the heterozygous tumors, or activated telomere lengthening in the homozygous background tumors. We measured the presence of the two main telomere lengthening pathways in the tumors by quantifying telomerase activity and the amount of C-circles, which are a measure for the activity of the alternative lengthening of telomeres (ALT) pathway [16]. We found that in only one out of nine homozygous tumors a telomere lengthening pathway was clearly activated (Figure 4b). Thus, activation of known mechanisms of telomere lengthening are likely not the cause of the stable genomes in the homozygous animals.

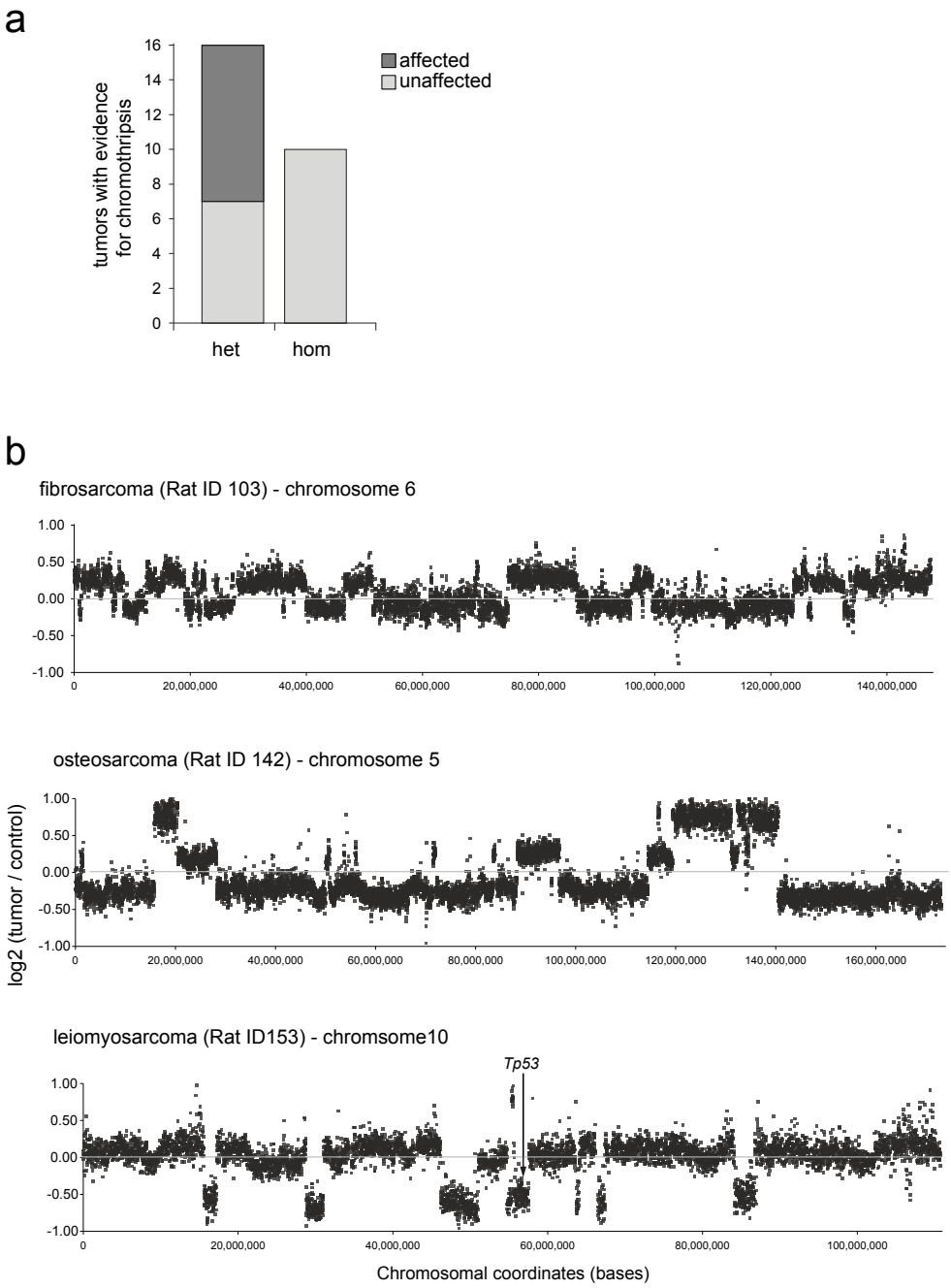


Figure 2. Tumors from homozygous *Tp53*^{C273X} mutant rats do not show any evidence for the occurrence of chromothripsis. (a) Chromothripsis signatures were counted in aCGH data from tumors from heterozygous animals (n=16) and tumors from homozygous animals (n=10). (b) Representative examples of chromothripsis in three different tumors from heterozygous animals. The lower panel shows LOH of the Tp53 locus as a result of chromothripsis.

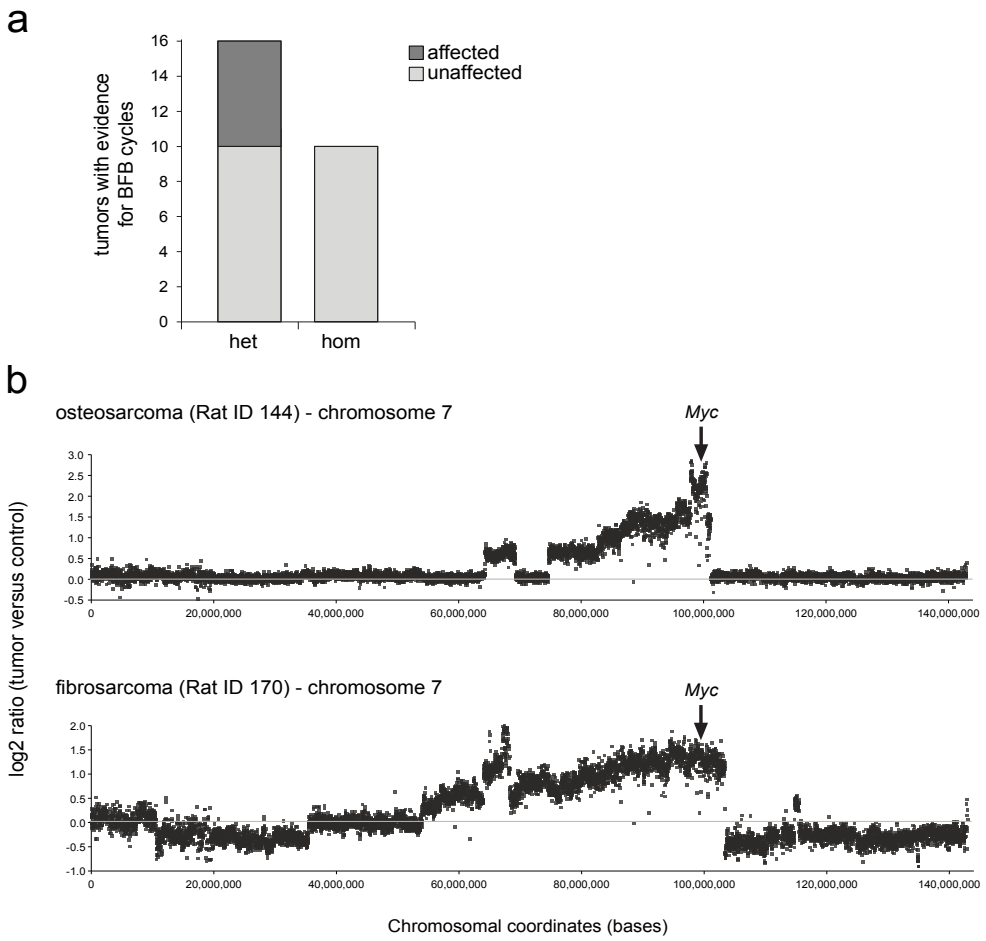


Figure 3. Tumors from homozygous *Tp53*^{C273X} mutant rats never show evidence for the occurrence of BFB cycles. (a) BFB cycle signatures were counted in aCGH data from tumors from heterozygous animals (n=16) and tumors from homozygous animals (n=10). (b) Representative examples of BFB cycles in two different tumors from heterozygous animals. Both panels show a gene amplification of the Myc oncogene on chromosome 7.

It is possible that these tumors have an alternative method to maintain telomere length, or that the difference in telomere length reflects telomeric shortening in the heterozygous animals.

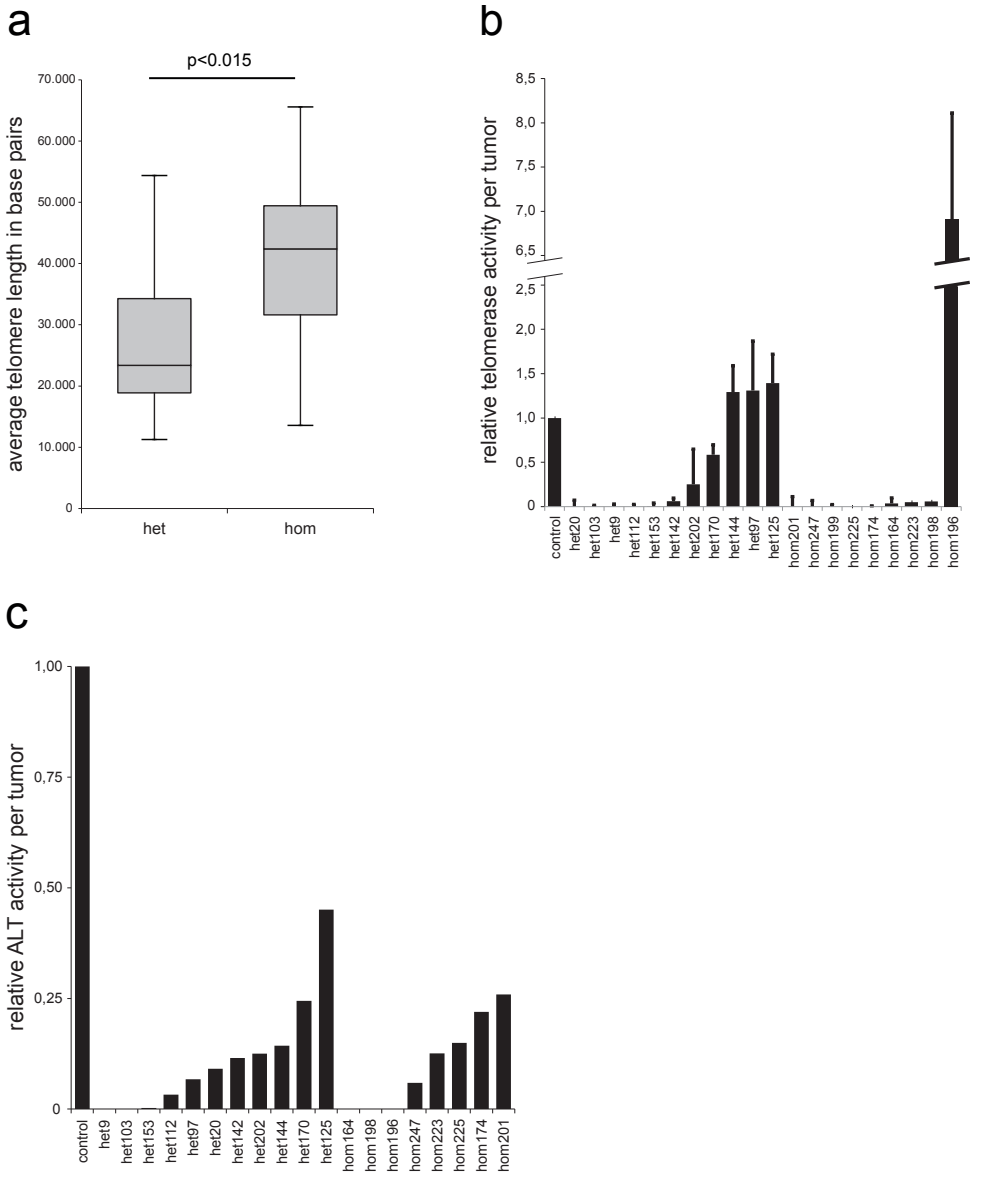


Figure 4. Tumors from homozygous *Tp53*^{C273X} mutant rats show significantly longer telomeres than tumors from heterozygous *Tp53*^{C273X} rats. (a) Length of an average telomere per tumor. Tumors from heterozygous animals (n=16) and tumors from homozygous animals (n=10) were compared using a Student's t-test. (b) Telomerase activity of tumor cell extracts, normalized for the telomerase-positive control cell line provided with the kit. Error bars represent s.e.m. (c) ALT activity of tumor cell extracts, normalized for the ALT-positive control cell line U-2 OS.

In summary we report here that the tumors in p53 heterozygous and null rats are different in many aspects. As was seen before [8] the tumor spectrum is different. Homozygous animals mainly develop hemangiosarcomas between 2 and 5 months of age, whereas heterozygous animals mainly develop osteosarcomas between 8 and 16 months of age. Our data show that this early occurrence of these tumors is not due to increased genomic instability. On the contrary, tumors from homozygous animals contain a low number of aneuploidies and CNV affected base pairs. Furthermore these tumors do not display any complex structural aberrations such as chromothripsis and BFB cycles. The genomes were particularly stable in the hemangiosarcomas. In this respect, it is relevant to look at the cell type of origin of hemangiosarcomas: the hemangioblasts [17-19]. Hemangioblasts are progeny from the bone-marrow-derived hematopoietic stem cells (HSCs). Stem cells, like HSCs, are known to actively maintain their telomeres, since they continuously produce progeny and thus cycle throughout life. In p53 null mice, HSCs expand and proliferate greatly and their apoptotic potential is reduced [20, 21]. The high proliferation rate in p53 null animals in combination with the telomere maintenance potential may explain the observed hemangiosarcoma tumors with a stable genome.

Conclusion

Overall our results underscore the principle of common sequence of oncogenic events in tumorigenesis [22]. The data show that the mere loss of functional p53 is not sufficient to induce large-scale genomic instability, thereby challenging the current p53-dogma.

Supplemental information

All supplemental files in this chapter can be downloaded from http://www.hubrecht.eu/research/cuppen/hermsen_thesis.html

References

1. Chang F, Syrjanen S, Tervahauta A, Syrjanen K: Tumourigenesis associated with the p53 tumour suppressor gene. Br J Cancer 1993, 68(4):653-661.

2. Hollstein M, Sidransky D, Vogelstein B, Harris CC: p53 mutations in human cancers. Science 1991, 253(5015):49-53.

3. Malkin D, Li FP, Strong LC, Fraumeni JF, Jr., Nelson CE, Kim DH, Kassel J, Gryka MA, Bischoff FZ, Tainsky MA et al: Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 1990, 250(4985):1233-1238.

4. Rausch T, Jones DT, Zapatka M, Stutz AM, Zichner T, Weischenfeldt J, Jager N, Remke M, Shih D, Northcott PA et al: Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell 2012, 148(1-2):59-71.

5. 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(7383):53-58.

6. Artandi SE, Chang S, Lee SL, Alson S, Gottlieb GJ, Chin L, DePinho RA: Telomere dysfunction promotes non-reciprocal translocations and epithelial cancers in mice. Nature 2000, 406(6796):641-645.

7. Chang S, Khoo CM, Naylor ML, Maser RS, DePinho RA: Telomere-based crisis: functional differences between telomerase activation and ALT in tumor progression. Genes Dev 2003, 17(1):88-100.

8. 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(4):1616-1622.

9. van Boxtel R, Toonen PW, Verheul M, van Roekel HS, Nijman IJ, Guryev V, Cuppen E: Improved generation of rat gene knockouts by target-selected mutagenesis in mismatch repair-deficient animals. BMC genomics 2008, 9:460.

10. van Heesch S, Mokry M, Boskova V, Junker W, Mehon R, Toonen P, de Bruijn E, Shull JD, Aitman TJ, Cuppen E et al: Systematic biases in DNA copy number originate from isolation procedures. Genome Biol 2013, 14(4):R33.

11. Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009, 25(14):1754-1760.

12. Lau LM, Dagg RA, Henson JD, Au AY, Royds JA, Reddel RR: Detection of alternative lengthening of telomeres by telomere quantitative PCR. Nucleic Acids Res 2013, 41(2):e34.

13. Overholtzer M, Rao PH, Favis R, Lu XY, Elowitz MB, Barany F, Ladanyi M, Gorlick R, Levine AJ: The presence of p53 mutations in human osteosarcomas correlates with high levels of genomic instability. Proc Natl Acad Sci U S A 2003, 100(20):11547-11552.

14. Baumhoer D, Gunawan B, Becker H, Fuzesi L: Comparative genomic hybridization in four angiosarcomas of the female breast. Gynecol Oncol 2005, 97(2):348-352.

15. Kitada K, Aikawa S, Aida S: Alu-Alu fusion sequences identified at junction sites of copy number amplified regions in cancer cell lines. Cytogenet Genome Res 2013, 139(1):1-8.

16. Henson JD, Cao Y, Huschtscha LI, Chang AC, Au AY, Pickett HA, Reddel RR: DNA C-circles are specific and quantifiable markers of alternative-lengthening-of-telomeres activity. Nat Biotechnol 2009, 27(12):1181-1185.

17. Gorden BH, Kim JH, Sarver AL, Frantz AM, Breen M, Lindblad-Toh K, O'Brien TD, Sharkey LC, Modiano JF, Dickerson EB: Identification of Three Molecular and Functional Subtypes in Canine Hemangiosarcoma through Gene Expression Profiling and Progenitor Cell Characterization. Am J Pathol 2014.

18. Lamerato-Kozicki AR, Helm KM, Jubala CM, Cutter GC, Modiano JF: Canine hemangiosarcoma originates from hematopoietic precursors with potential for endothelial differentiation. Exp Hematol 2006, 34(7):870-878.

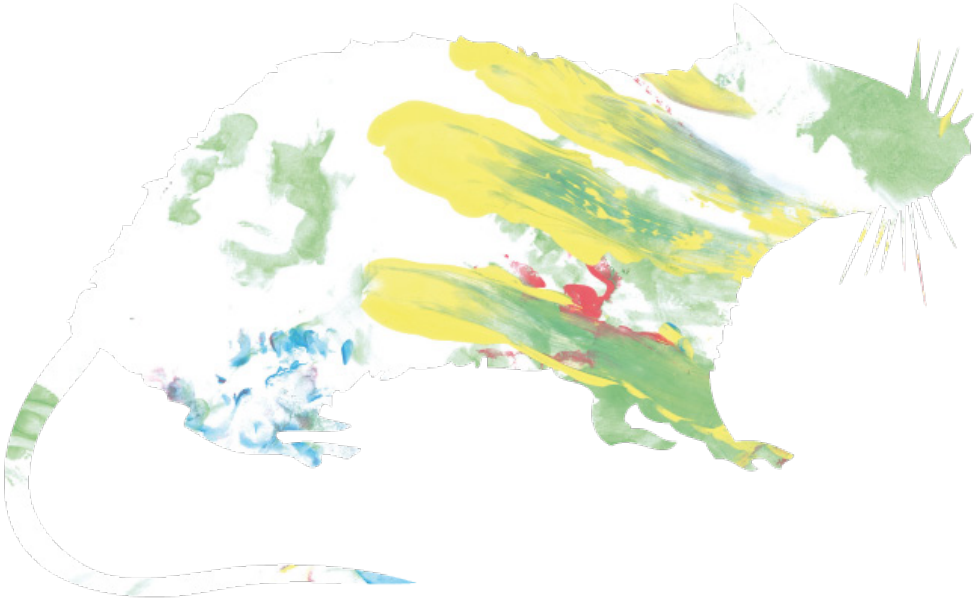
19. Liu L, Kakiuchi-Kiyota S, Arnold LL, Johansson SL, Wert D, Cohen SM: Pathogenesis of human hemangiosarcomas and hemangiomas. Hum Pathol 2013, 44(10):2302-2311.

20. Milyavsky M, Gan OI, Trottier M, Komosa M, Tabach O, Notta F, Lechman E, Hermans KG, Eppert K, Konovalova Z et al: A distinctive DNA damage response in human hematopoietic stem cells reveals an apoptosis-independent role for p53 in self-renewal. Cell Stem Cell 2010, 7(2):186-197.

21. TeKippe M, Harrison DE, Chen J: Expansion of hematopoietic stem cell phenotype and activity in Trp53-null mice. Exp Hematol 2003, 31(6):521-527.

22. Vogelstein B, Kinzler KW: Cancer genes and the pathways they control. Nat Med 2004, 10(8):789-799.

6



Chapter 6

General discussion

Since its introduction nearly ten years ago, next-generation sequencing has had a profound impact on the rat functional genomics field (reviewed in **Chapter 1**). The use of NGS further accelerated following the publication of the first rat reference sequence. Not only did whole genome sequencing become a standard lab technique, but countless other NGS-based applications assaying different cellular entities, like transcriptome or chromatin status, were also developed. Research, but also diagnostics, was heavily revolutionized by the introduction of NGS. New assays are being created and further developed which have the potential for broad application in general healthcare.

In this thesis I described work that assesses the use of NGS based applications in rat functional genomic research as a tool to identify and molecularly characterize genomic variants in complex traits. In the **General discussion** I will elaborate further on my findings and future directions of this work. Furthermore I will discuss the implications of my work and NGS in general on current healthcare and society, because discussion on the impact of NGS on society is often neglected.

Data integration and visualization

One of the main challenges in research areas that utilize NGS applications will be the direct integration of data. This includes integration of one type of data between different studies and secondly between different data modalities. For whole genome sequencing data this will require mining from public repositories and mapping on a single version of the rat reference genome, after which simultaneous variant calling can be done. In **Chapter 2** I use this approach to identify substrain variants, but in general this will be an approach that will be continually repeated when an improved reference genome or variant caller is published. This process is essential to retain adequate genotype information from all sequenced rat strains at once. In this way similarities and differences between strains are easily detectable and strains become more comparable. Thus handling the integration of a single data type is relatively straightforward. However, the integration of different data modalities will be more challenging. In **Chapter 4** I describe the integration of genomic and transcriptomic data in combination with epigenetic marks and 3D chromatin organization. In this work I show that integration is indeed possible in a correlative fashion. The next step would be data-based model building and testing of hypotheses. However, NGS data, as all techniques, comes with its technical

shortcomings which influence the data generation. Incompleteness of NGS data sets in combination with differing (in)completeness between the NGS data types are currently limiting model building. Also other interdisciplinary methods will have to be developed in which biologists and mathematicians will cooperate to integrate biological premises and data modalities in mathematical models. A first impressive step was taken by Karr *et al*, which modeled and predicted a whole-cell life cycle of the human pathogen Mycoplasma genitalium [1].

Dissecting complex disease and annotation of the noncoding genome

Another challenge in the quest for one or multiple causal genomic loci underlying a complex disease is the often polygenic nature of these traits. In addition, the etiology of these diseases is not fully understood.

Studying human complex disease in the rat model system allows the dissection of these traits, since these models resemble the human disease situation. However, in **Chapter 5** we make observations that do not fit the human or mouse etiology, because we find that tumors from *Tp53* null rats display stable genomes in contrast with unstable tumors in mouse. Furthermore, only Li-Fraumeni patients come closest to the inborn p53 null status, since no patients have been reported to be germline *TP53* null. The observation that tumors of homozygous *Tp53* null rats display stable genomes challenges the current p53 dogma and can give us insights to initiate further research in an attempt to understand the complex disease etiology of cancer.

In order to identify genetic variants that underlie complex disease, **Chapter 3** describes the identification of 355 QTLs for 122 phenotypes in the heterogeneous stock cross. This unique and massive cross builds on the availability of the full genome sequence of the eight progenitor strains, allowing the genomes of their 1407 descendants to be imputed. The semi-outbred nature of this cross results in frequent crossing overs that in turn result in relatively small QTLs with a median size of 4.4 Mb. In this way the search space for phenotype-underlying causal loci is limited. However, the majority of detected QTLs still contain multiple genes of which most do not contain a perturbing variant. The pivotal role of the noncoding genome is also apparent in the human GWAS field, where the vast majority of detected disease-associated variants reside in the

noncoding part of the genome [2-4]. In **Chapter 4** I show important steps in unraveling the complex regulatory role of the noncoding genome. This is exemplified by a deletion that deregulates the expression of the *Pkhd1*, a gene that is more than 150 Kb away. This and other findings force us to reconsider our variant annotation and the valuation of variants in the noncoding genome. This will be a challenge since multi-level data is required for proper annotation, which will not be available in all cases. Furthermore, the continued investment in basic knowledge on the noncoding genome, such as the ENCODE project, is essential to get more insight in this matter. The start of modENCODE for model organisms will contribute to this knowledge and it will be extremely beneficial for rat studies when we will see rat being subjected to such an approach.

My work from public perspective

The research described in this thesis is a basis for further fundamental research and reports multiple genome-wide resources, which have the potential to change current healthcare. The current knowledge needed for the interpretation of all this data is not yet sufficient, but the contribution of my work is an important step in the dissection of complex disease in man. One contribution is by presenting a comprehensive database containing genomic variants and its utility (**Chapter 2 & 3**). This resource contains potential candidate variants that underlie complex traits. The next step, the pinpointing of these variants, will be a larger challenge for the near future. Secondly, my work contributes to the understanding of human biology and disease by gaining insight in the type of consequence a genomic variant can have and in how these variants can interfere in molecular processes and thus disease etiology (**Chapter 4 and 5**). Having said this, I understand that these statements are still quite broad and not very concrete. Yet the shift towards a direct applicable result from scientific research is something that is currently intensively stimulated both by government and funding parties. In a way this is understandable, but still the role of fundamental research is often underestimated. Without gaining 'knowledge for knowledge's sake' and opportunism, scientific progress will stall instantaneously. Along this same perspective I also would like to discuss the use of laboratory animals in the formation of this thesis. In the current dogma the use of animals in science is important for better understanding in biology. Experiments that deviate from the conventional scientific route or employ 'scientific creativity' typically have a 'high risk – high gain' characteristic. Although the policy and feeling

to minimize animal usage is understandable from a public perspective, 'creative science' has high value for scientific progress and thus for society in general. To gain some perspective in this discussion: the work presented in this thesis used a total of 1928 laboratory rats. I realize that the balance between animal usage and 'creative science' is delicate and this topic can benefit from more discussion in scientific research groups.

The use of "omics" data in the public playing field

The impact of next-generation sequencing on society has just started. The costs will drop and the one thousand dollar genome will soon be within reach. Introduction of NGS applications in current healthcare in the Netherlands has started in the last few years. The classic invasive amniocentesis technique was replaced by the non-invasive prenatal test (NIPT), by which fetal DNA in the maternal blood is screened for trisomy 13, 18 and 21 and sex. This current application shows that NGS is no longer an assay restricted to fundamental researchers but that it has big implications for society. Although the techniques were introduced almost a decade ago, the public genetic awareness of the potential and the impact of NGS is still minimal. In this light I initiated a project to challenge people to think about the possible implications of NGS by confronting them with the question: With whom would you share the sequence of your own DNA? Together with my team we conceptually developed the idea of a DNA predictor: an interactive display in the D&A popup stores funded by the Netherlands Genomics Initiative [5], which visited 13 cities in the Netherlands. On integrated touchscreens in this interactive display participants answered 10 questions about their phenotype, such as "Do you have freckles?", and other personal questions. After the questions, a personalized DNA code was generated based on the known SNPs underlying these phenotypes. Participants could then share this on Facebook or send it to their own email address, representing the 'public' and 'private' sharing options. In addition, participants received a personalized T-shirt with their own DNA-code. Altogether 4,200 T-shirts were handed out to more than 10,000 store visitors. Six percent (146 people) of the DNA predictor participants shared their DNA on Facebook of the 2277 participants that fully completed the 'test'. Although people were asked to discuss the subject of sharing your own DNA with others, only a small percentage in the end shared their DNA sequence on Facebook. This might indicate that people are still hesitant to share such personal information, which I believe is a healthy characteristic in this time of

rapidly changing technology and access to information. The ultimate role of NGS in our society is a topic for wide speculations. The idea of a genetic passport with interpretable data for the doctor, insurance company and employer is still a thought experiment, but could very well be reality in the coming decade. As it has done in the rat research field, NGS is likely to also revolutionize daily human health care practice and may thus also affect your life. It is not a question of how large the role of NGS will be, but how you as a citizen will cope with it. So be prepared, discuss and be proactive!

References

1. Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival, B., Jr., Assad-Garcia, N., Glass, J.I. and Covert, M.W. (2012) A whole-cell computational model predicts phenotype from genotype. Cell, 150, 389-401.
2. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America, 106, 9362-9367.
3. Kumar, V., Wijmenga, C., and Withoff, S. (2012) From genome-wide association studies to disease mechanisms: celiac disease as a model for autoimmune diseases. Seminars in immunopathology 34, 567-580.
4. Pennisi, E. (2011) The Biology of Genomes. Disease risk links to gene regulation. Science, 332, 1031.
5. Netherlands Genomics Initiative, http://www.genomics.nl/. 2008-2012.

6

&

Addendum

Nederlandse samenvatting
Dankwoord
List of publications
Curriculum vitae

Nederlandse Samenvatting

DNA is de erfelijke informatie waarin de bouwtekening te vinden is van je lichaam. Elke cel in je lichaam bevat DNA (je 'genoom') en heeft dit nodig om goed te kunnen functioneren. Van het DNA worden kleine delen (genen) afgeschreven tot eiwitten. Deze eiwitten zijn de 'werkpaarden' van een cel en zijn betrokken bij alle belangrijke celprocessen. Ook al is meer dan 99,9% van het DNA identiek in alle mensen, de 0,1% die wèl verschilt speelt een belangrijke rol. Deze verschillen, ook wel 'DNA varianten' genoemd, zijn allereerst verantwoordelijk voor verschillen tussen mensen in uiterlijke kenmerken, zoals oog- en haarkleur. Daarnaast hebben ze ook invloed op 'interne' zaken zoals de vochtigheid van je oorsmeer en zelfs vatbaarheid voor ziekten. Zo zijn er ziekten waarbij één duidelijke DNA variant de oorzaak is, zoals taaislijmziekte. Maar daarnaast zijn er ook ziekten (eigenschappen) die beïnvloed worden door veel verschillende DNA varianten tegelijkertijd, zoals hoge bloeddruk. Deze laatste eigenschappen zijn zogenaamd complex. De rat als modeldier voor de mens wordt al 150 jaar ingezet om meer te leren over deze complexe eigenschappen.

DNA varianten kun je detecteren als je het DNA van een organisme afleest (DNA sequencing). Door de komst van een nieuwe manier van DNA aflezen vanaf 2005, de zogenaamde 'next-generation' sequencing (NGS), is het mogelijk geworden om snel en nauwkeurig het hele genoom van een mens of een modelorganisme af te lezen (een overzicht van deze ontwikkeling is te vinden in **Hoofdstuk 1**). Hierdoor is er in korte tijd al een grote hoeveelheid genetische informatie gegenereerd, die gebruikt kan worden om DNA varianten (genotype) te koppelen aan complexe eigenschappen (fenotype). Deze koppeling is lastig te bestuderen in de mens doordat deze altijd onderhevig is aan variabele milieu-invloeden. Tevens is het menselijk genoom lastiger te interpreteren dan het genoom van een inteelt modeldier, zoals de rat. Daarom worden modeldieren als de rat of de muis vaak ingezet voor onderzoek naar de relatie tussen genotype en fenotype.

In dit proefschrift bestudeer ik de onderliggende DNA varianten van complexe eigenschappen in de rat als modelorganisme voor de mens. Ik probeer aanwijzingen te vinden die ons meer kunnen vertellen over de relatie tussen genotype en fenotype. Een belangrijke systeem op deze

relatie in te bestuderen zijn ratten lijnen. Deze lijnen zijn speciaal zo gefokt zodat ze een constant genotype en fenotype hebben. Allereerst beschrijf ik in **Hoofdstuk 2** de inventarisatie van DNA varianten in 40 ratten lijnen. Deze lijnen worden over de hele wereld gebruikt in onderzoek naar complexe eigenschappen en ziekten, waaronder diabetes en hoge bloeddruk. In dit hoofdstuk laten we zien waar in het genoom deze varianten liggen en wat precies de verschillen zijn tussen de 40 ratten lijnen.

Van deze catalogus van varianten maken we in **Hoofdstuk 3** gebruik, door deze te koppelen aan 122 fenotypes in 1408 speciaal gefokte ratten. Deze ratten stammen allemaal af van 8 geselecteerde ratten lijnen af, waarvan we de DNA varianten kennen. In totaal vinden we 355 koppelingen tussen het genotype en het fenotype. Deze koppelingen vormen een basis voor veel verder onderzoek, aangezien elke koppeling apart in detail bestudeerd moet worden om de achterliggende mechanismen volledig te begrijpen.

Om meer inzicht te krijgen in deze mechanismen, bestuderen we in **Hoofdstuk 4** de gevolgen van DNA varianten en meer specifiek: DNA varianten die niet in een gen liggen. De interpretatie van de gevolgen van dit soort varianten is tot nu toe lastig en wij proberen hier meer inzicht in te krijgen. We laten zien op welke manier DNA varianten buiten genen invloed kunnen hebben op het fenotype. Ze kunnen bijvoorbeeld het gebruik van een gen reguleren, door het gen vaker of juist minder vaak te gebruiken. Hierdoor wordt er meer of minder eiwit afgeschreven, en verandert het functioneren van een cel of zelfs (een deel van) het hele lichaam. Ook laten we zien dat DNA varianten lijken te kunnen zorgen voor een andere manier van DNA vouwing, wat uiteindelijk ook kan bijdragen aan verschillen in het afschrijven van een eiwit. Doordat, per cel, het 2 meter lange DNA in een celkern van 0,01 millimeter moet worden ‘gepropt’, moeten er keuzes gemaakt worden tussen welke delen beschikbaar moeten blijven voor bijvoorbeeld het gebruik van genen, en welke delen minder goed beschikbaar hoeven zijn. Wij laten in **Hoofdstuk 4** een mooi voorbeeld zien van een DNA variant die invloed lijkt te hebben op deze vouwing en daardoor dus gen-gebruik op een indirecte manier zou kunnen beïnvloeden.

Naast het begrijpen van de koppeling en functie van DNA varianten in complexe eigenschappen, is het ook van belang om meer te weten over

de oorzaken en mechanismen van de complexe eigenschappen of ziekten zelf. In **Hoofdstuk 5** beschrijf ik de gedetailleerde karakterisatie van het genoom van een aantal tumoren afkomstig uit een speciaal hiervoor gefokt rat model. In veel tumoren ziet het genoom in de tumor er anders uit dan in de rest van het lichaam: er zijn allerlei veranderingen opgetreden in het DNA die het deels mogelijk maken voor de tumor om uit te groeien. Onderzoek tot nu toe heeft laten zien dat een tumor zogenaamd 'genomisch instabiel' is. In ons rat model zien we verassend genoeg dat sommige van deze tumoren juist een heel stabiel genoom hebben. Dit is een observatie die ingaat tegen het huidige dogma. Deze nieuwe informatie is interessant om aanwijzingen te krijgen voor vervolgonderzoek, hoe kanker in dit rat model ons meer kan vertellen over de oorzaak en mechanismen achter kanker in het algemeen.

In mijn proefschrift heb ik de nieuwe 'next-generation' manier van sequencing al veel ingezet voor fundamenteel onderzoek. Echter, de rol van NGS in de huidige gezondheidszorg begint ook steeds belangrijker te worden. Er wordt telkens naar meer toepassing van deze techniek gezocht en het zal niet lang duren voordat ook buiten de gezondheidszorg de impact van NGS in de maatschappij voelbaar is door ons allemaal. Zo wordt er nu al gebruik gemaakt van NGS om op een niet-invasieve manier het syndroom van Down, Patau en Edwards te diagnosticeren op basis van rondzwevend foetaal DNA in de bloedbaan van de moeder. In theorie is het zelfs al mogelijk om vroeg in de ontwikkeling alle DNA varianten in dit foetale DNA in kaart te brengen. De vraag is dan natuurlijk of de maatschappij dit wel wil? In **Hoofdstuk 6** bediscussieer ik, naast de bevindingen in mijn proefschrift, ook deze invloed van NGS op de maatschappij; nu en in de toekomst.

Dankwoord

Het boekje is af. BAM!

Vier jaar geleden begon ik aan een mooi avontuur waarin ik veel geleerd heb. In deze tijd ben ik op vele manieren geholpen, wat uiteindelijk geleid heeft tot de totstandkoming van dit proefschrift. Daarom wil deze mensen bedanken voor hun steun en support.

Allereerst natuurlijk Edwin, mijn mentor. Bedankt voor de mogelijkheid om mijn PhD in je lab te mogen doen. In deze tijd heb je me de vrijheid gegeven om mijn PhD op mijn eigen manier te mogen invullen en die kans heb ik met beide handen aangegrepen. Naast dat je me leerde om gedegen onderzoek te doen en te rapporteren, kon ik ook buiten het onderzoek mezelf ontplooien door o.a. deelname aan de Academische Jaarprijs en uiteindelijk het ontwikkelen van de DNA Predictor met een aantal collega's. Ik heb dit als zeer waardevol ervaren en vind het fijn dat je me die kans hebt gegeven.

Marieke, Ewart (K.) en Victor, altijd kon ik met vragen bij jullie terecht, waarvoor erg bedankt. Jullie hebben mede gevormd tot een zelfbewuste wetenschapper op jullie eigen manier. Die variëteit in samenwerking en feedback vond ik erg prettig.

&

Mijn paranimfen: Pim en Sebas. Allereerst Pim, ik heb vele mooie herinneringen van het werk samen in de (ratten)stal (maar zeker ook daarbuiten) en jouw enthousiasme werkt altijd aanstekelijk, ook op momenten dat ik het licht even niet zag. Hopelijk tot snel in (bijvoorbeeld) het Máximapark. Sebas, samen gestart als PhD students en nu samen klaar. Veel van elkaar geleerd en veel mooie momenten in diverse steden wereldwijd, vaak gepaard met de nodige biertjes. Ik ga je missen man!

Dan natuurlijk mijn lieve collega's van de Cuppengroup: Het was een match-made-in-heaven; na het zien van de bierkoelkast tijdens mijn tweede sollicitatiegesprek dacht ik: "Hier ga ik geen spijt van krijgen!". Monique, Frans Paul, Henk, Ewart (dB.), Maarten, Wensi, Wim, Lisanne, Sander, Mark, Nico, Esther. Maar ook de (ex)Cuppen leden van het UMC: Terry, Mark, Marlous, Mirjam, Nayia, Kirsten, Ivo, Ies, Pjotr, Martin, Karen,

Marco, Nicolle, Petra, Glen, Glenn en Magdalena. Jullie hebben mijn tijd op het lab tot een waar feest gemaakt.

Op de AIO kamer allereerst dank voor het uitwisselen van kennis, kunde, leed en plat vermaak Michal en Jos. Maar ook Francis en Myrthe, de nieuwe frisse garde, bedankt voor jullie inhoudelijke en sociale steun in de laatste fase van mijn promotie.

Mijn enige 'eigen' studente Maryvonne en mijn surrogaat studenten Kim, Robin en Rutger wil ik ook bedanken. Ik vond het erg leuk jullie, gevraagd en ongevraagd, te mogen begeleiden. Mede door jullie heb mijn link met onderwijs en begeleiding gehouden en keer ik nu, na mij promotie, daar weer naar terug.

A special thanks to all collaborators for the various contributions for all chapters in my thesis: From Utrecht, the group of Alain de Bruin, including Sameh and Raoul for Chapter 5; from all over Europe, the members of the EURATRANS consortium for Chapter 2, 3 and 4; and from the United States, the group of Boris Tabakoff for Chapter 2.

Uiteraard wil ik alle Hubrechters bedanken voor een leuke tijd, de afgelopen vier jaar. Met een speciale dank aan onze derde-verdieping-buddy-groep de Creyghtons. Ook wil ik de dierverzorgers bedanken voor de goede zorgen voor onze ratten.

Naast professionele steun ben ik privé natuurlijk ook bijgestaan door de dinsdag Supernerden vrienden, m'n (schoon)familie, m'n broer en schoonzus en m'n (schoon)ouders. Fijn dat onze ontspanningsmomenten altijd in goede balans waren met mijn promotiewerkzaamheden.

En tot slot, mijn eigen gezin, de motor achter mijn motivatie. Lieve Tom, mijn laatste promotiejaar was erg turbulent en nog steeds word ik gevuld met energie als ik bij jou ben! Lieve Merel, ik vond het fijn dat je me altijd hebt scherp gehouden, ook al vroeg ik daar niet altijd om. Mede door jou heb ik dit tot een goed einde weten te brengen. En ik hoop voor volgende 'projecten' je weer als inspiratie te mogen gebruiken!

&

List of publications

Baud A, Guryev V, Hummel O, Johannesson M, Hermsen R, Stridh P, Graham D, McBride MW, Foroud T, Calderari S, Diez M, Ockinger J, Beyeen AD, Gillett A, Abdelmagid N, Guerreiro-Cacais AO, Jagodic M, Tuncel J, Norin U, Beattie E, Huynh N, Miller WH, Koller DL, Alam I, Falak S, Osborne- Pellegrin M, Martinez-Membrives E, Canete T, Blazquez G, Vicens-Costa E, Mont-Cardona C, Diaz-Moran S, Tobena A, Zelenika D, Saar K, Patone G, Bauerfeind A, Bihoreau MT, Heinig M, Lee YA, Rintisch C, Schulz H, Wheeler DA, Worley KC, Muzny DM, Gibbs RA, Lathrop M, Lansu N, Toonen P, Ruzius FP, de Bruijn E, Hauser H, Adams DJ, Keane T, Atanur SS, Aitman TJ, Flicek P, Malinauskas T, Jones EY, Ekman D, Lopez-Aumatell R, Dominiczak AF, Holmdahl R, Olsson T, Gauguier D, Hubner N, Fernandez- Teruel A, Cuppen E, Mott R, Flint J: Genomes and phenomes of a population of outbred rats and its progenitors. **Scientific Data** 2014, 1:140011

Baud A, Hermsen R, Guryev V, Stridh P, Graham D, McBride MW, Foroud T, Calderari S, Diez M, Ockinger J, Beyeen AD, Gillett A, Abdelmagid N, Guerreiro-Cacais AO, Jagodic M, Tuncel J, Norin U, Beattie E, Huynh N, Miller WH, Koller DL, Alam I, Falak S, Osborne- Pellegrin M, Martinez-Membrives E, Canete T, Blazquez G, Vicens-Costa E, Mont-Cardona C, Diaz-Moran S, Tobena A, Hummel O, Zelenika D, Saar K, Patone G, Bauerfeind A, Bihoreau MT, Heinig M, Lee YA, Rintisch C, Schulz H, Wheeler DA, Worley KC, Muzny DM, Gibbs RA, Lathrop M, Lansu N, Toonen P, Ruzius FP, de Bruijn E, Hauser H, Adams DJ, Keane T, Atanur SS, Aitman TJ, Flicek P, Malinauskas T, Jones EY, Ekman D, Lopez-Aumatell R, Dominiczak AF, Johannesson M, Holmdahl R, Olsson T, Gauguier D, Hubner N, Fernandez-Teruel A, Cuppen E, Mott R, Flint J: Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. **Nat Genet** 2013 45:767-775.

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.

Hermsen R, de Ligt J, Spee W, Blokzijl F, Schafer S, Adami E, Boymans S, Flink S, van Boxtel R, van der Weide RH, Aitman T, Hubner N, Simonis M, Tabakoff B, Guryev V, Cuppen E: Genomic landscape of rat strain and substrain variation. *Submitted*

van Heesch S*, Hermsen R*, Lansu N, de Luca K, Spee W, Boymans S, de Bruijn E, Toonen P, Verheul M, Thybert D, Flicek P, de Laat W, Cuppen E, Simonis M: Multilevel effects of non-coding genetic variation on regulatory elements and chromatin organization. *Submitted* **Contributed equally*

Hermsen R, Toonen R, Hassan S, Kuiper R, Kuijk E, de Bruin A, Cuppen E, Simonis M: Lack of major genome instability in tumors of p53 null rats. *Submitted*

van der Weide RH, Simonis M, Hermsen R, Cuppen R: The scrapheap challenge: extracting relevant biological information, including strain-specific genomic segments, from unmapped NGS-reads. *Manuscript in preparation*

Curriculum Vitae

Roel Hermsen was born on January 23, 1984 in Huissen, The Netherlands. In 2003 he graduated from the Stedelijk Gymnasium Arnhem after which he started his study Medical Biology at the Radboud University Nijmegen. He performed his first master internship at the department General Internal Medicine at the Radboud University Medical Centre Nijmegen under supervision of dr. Marcel van Deuren. Roel performed his second internship at the department of Infectious Diseases and Immunology at the Faculty of Veterinary Medicine at the Utrecht University under the supervision of dr. Ildiko van Rhijn. In 2009 he graduated, receiving a M.Sc. degree in Medical Biology after which he started the post-academic master Teaching Biology for secondary education. In 2010 he graduated receiving a M.Sc. degree. In August that same year he began working as a PhD candidate in the Genome Biology group of Prof.dr. Edwin Cuppen at the Hubrecht Institute in Utrecht. The results of this work are described in this thesis. In September 2014, Roel started as a lecturer Molecular Biology at the University of Applied Sciences Leiden.