Dissertation

submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of Doctor of Natural Sciences

Presented by Jonathan Landry
born in Lyon, France
Oral examination: 06/12/2012

The genomic and transcriptomic landscape of HeLa cells

Referees: Dr. Jan Ellenberg
Prof. Dr. Stefan Wölfl

Abstract

HeLa is and always has been the most widely used model cell line for studying human cellular and molecular biology, and has thus defined our current understanding of human cells. However, no genomic or transcriptomic references exist for this cell line, and studies thus far have relied on the undoubtedly different human reference genome. This knowledge is essential to guide the increasing number of molecular and genetic studies being done in this cell line, both for design and interpretation. During my doctoral work, I performed deep DNA and RNA sequencing of a HeLa-Kyoto cell line and analyzed its mutational portfolio and gene expression profile. We generated a comprehensive description of the genomic and transcriptomic landscape of this cell line. A segmentation of the genome according to copy number revealed an astonishingly high level of aneuploidy and numerous large structural variations at unprecedented resolution. Remnants of catastrophic chromosome shattering, known as chromothripsis, were evident. Comparison of the HeLa expression profile to the physiological range of human gene expression reveals that several pathways, including cell cycle and DNA repair mechanisms, are 'abnormally' expressed. These results provide the first detailed account of the extent of variations in the HeLa genome, yielding insight into their impact on gene expression, cellular function, and their origins during the evolution of this cell line. This study demonstrates the importance of accounting for the strikingly aberrant characteristics of HeLa cells when designing and interpreting experiments. The important resources provided will help to reevaluate the way HeLa is used to model human biology.

Zusammenfassung

HeLa ist momentan die in der menschlichen Zell- und Molekularbiologie am weitesten verbreitete Modell-Zelllinie und hat unser Verständnis menschlicher Zellen maßgeblich beeinflusst. Da weder ein Referenzgenom noch ein Referenztranskriptom für HeLa verfügbar ist, verwendeten bisherige Studien das zweifelsohne unterschiedliche humane Referenzgenom zur Datenanalyse.

Ein verfügbares Referenzgenom für HeLa ist essentiell um das Design und die Interpretation einer wachsenden Anzahl molekularer und genetischer Studien an dieser Zellline zu ermöglichen. In meiner Doktorarbeit habe ich DNA und RNA einer HeLa-Kyoto Zellline sequenziert und ihr Mutationsportfolio und Genexpressionsprofil analysiert, sowie eine umfassende Charakterisierung der genomischen und transkriptomischen Landschaft dieser Zellline erstellt.

Die Segmentierung des Genoms nach “copy number” enthüllt ein erstaunlich hohes Maß an Aneuploidie sowie zahlreiche große strukturelle Variationen mit einer bisher nicht verfügbaren Auflösung. Die Spuren von katastrophalen chromosomalen Umlagerungsereignissen, bekannt als Chromothripsis, sind evident in dieser Segmentierung. Der Vergleich des Expressionsprofils von HeLa mit dem physiologischen Spektrum menschlicher Genexpression offenbart mehrere Signalwege, unter anderem Zellzyklus und DNA Reperaturmechanismen, deren Expressionslevel stark von diesem Spektrum abweichen.

Diese Resultate liefern den ersten detaillierten Katalog genetischer Variationen in HeLa, und geben Einsicht in deren Einfluss auf Genexpression, zellulare Funktionen sowie die evolutionäre Herkunft während der Evolution dieser Zellinie. Diese Studie verdeutlicht, wie wichtig es ist, die abweichenden Charakteristika von HeLa beim Design und der Interpretation von Experimenten zu Berücksichtigen. Die von uns zur Verfügung gestellten Ressourcen werden dabei helfen, die Art und Weise in der HeLa als Modell humaner Biologie verwendet wird, neu zu bewerten.

Acknowledgements

I would like to start by thanking Lars for welcoming me into his lab during this PhD. Thank you for giving me the opportunity to work with you and your team for your availability, your enthusiasm, your vision, your passion for science and your support. I remember that you were referring to your relationship with Lee Hood as “standing on the shoulders of giants”. I feel really what it means now it has been a privilege to be part of your team. I would also like to thank the members of my thesis advisory committee, Jan Ellenberg, Jan Korbel, and Stefan Wölfl for all the helpful discussions and advice they have provided me with. I want to thank Stefan Wiemann for agreeing to be on my thesis defense committee.

I am also very grateful for the great collaborative atmosphere at EMBL with different groups and especially Jan Ellenberg’s, Jan Korbel’s and Wolfgang Huber’s group for all the fulfilling scientific discussions and to sharing your experience and expertise with me. Thanks to Faba, Jutta, Thomas and Jean-Karim for your feedback to grasp all the complexity of the HeLa cell line background and to communicate all your passion about your work. I also want to send many thanks to Tobias and Thomas for their help and their feedback on my work. I also want to mention that I feel very lucky to work closely with Wolfgang and his team. The quality of the science and feedback is always very appreciated and helpful. I would like to also to address a special thank to Paul, it has been a pleasure working with you. Thanks for your patience, your advice and all your help for the project. I hope that I did not bother you too much and too many times but I think it was worth it and that we built a nice story together, Danke!

I would like to address my gratitude to Anna Jauch, and her team, from the department of Human Genetics at Heidelberg University for sharing their expertise on cytogenetics with me.

To all the members of the Steinmetz group, thank you all for creating such an enjoyable daily atmosphere. I truly believe that this alchemy between all the lab members contributes to all the excellent work performed by the group. It is very important for me to mention all of you: Raeka, Sophie, Simon, Stephanie, Christophe, Sandra, Lin Gen, Ishaan, Aino, Petra, Eugenio, Aleksandra, Vicent, Xialong, Chao Tang, Manu, Wave, Stefan, Xiaobin, Judith, Zhenyu and Chenchen. Thanks to Manu, Nico, and Julien for all your help and discussion around the project and to Emilie and Raeka for all your time, your feedback on the writing of this document, a fantastic help.

During time before EMBL, I feel fortunate to cross the path of many people who give me encouragement and advice on my studies and career, I would like to thank them in Cambridge, Annabel Cook, Gillian Fraser, Christina Kahramanoglou, Nick Luscombe, Anabel Prieto, Aswin Seshasayee and Juanma Vaquerizas; in Paris, Ludovic Lacroix, Vladimir Lazar and Thomas Robert; in Montpellier, Pascale Chardenot, Michel Didier, Nadine Gonalons and Florence Oury-Donat. You help me to construct my professional path step by step.

To all my friends, it was great to have your support and you all made this experience very enjoyable. I will remember all the nice memories in Heidelberg and everywhere else.

Merci aux ESILiens: “parce que GBMA forever!”

Merci aux gones, toujours ressourçant de vous revoir lors de nos retours dans la ville des lumières.

Merci a toute ma famille pour votre soutien depuis toutes ces années. Ce travail vous est dédié et ceci n’aurait jamais été possible sans vous.

Loulou et Louloutte, ces 4 ans n’ont pas été seulement une expérience professionnelle enrichissante mais cela m’a aussi et surtout permis d’apprendre mon rôle et mes responsabilités de père de famille, je me sens heureux et accompli grâce a vous et aussi tellement chanceux de vous avoir !

Contents

List of Figures

List of Tables

List of Tables

1 Introduction

1.1 HeLa origins

1.1.1 Henrietta Lacks

The cells, that gave rise later to the HeLa cell line, were taken from a cervical cancer tumor of a patient called Henrietta Lacks. This 30-year-old woman was admitted in the Johns Hopkins Hospital in Baltimore, Maryland in 1951 for severe abdominal pain. The pain became intractable, even after several therapeutic attempts and she died of her cancer 2 months after her admittance to the hospital.
A few months earlier, she was diagnosed for cervical cancer subsequent to a cervix examination that revealed lesions of limited size [1]. A biopsy of the cervical tissue was performed and the pathology department diagnosed a epidermoid carcinoma, cervix uteri, spinal cell type with invasion of the stroma [2]. Even with radium and x-ray treatment during the following months, the cancer grew rapidly and aggressively and eventually caused her death. Her autopsy of the body confirmed the presence of different tumor nodules across different parts of the body with an enrichment of tumor in the pelvic area affecting the abdominal wall, the bladder, the kidney, the ureter, the cervix, the uterus, fallopian tubes, the ovaries and the vagina.

1.1.2 Cervical cancer

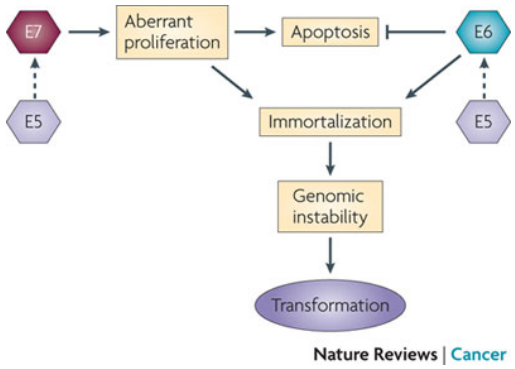
Cervical cancer accounts for 13% of all female cancers and represents the third most common cancer occurring in women and the seventh overall, with an estimated 530 000 new cases in 2008[1]. The highest incidence rates are in Eastern and Western Africa (30 per 100,000), Southern Africa (26.8 per 100,000), South-Central Asia (24.6 per 100,000), South America (23.9 per 100,000) and Middle Africa (23.0 per 100,000 respectively) and lowest are in Western Asia, Northern America and Australia/New Zealand (less than 6 per 100,00). Overall, the mortality: incidence ratio is 52%, and cervical cancer was responsible for 275 000 deaths in 2008.
In the 1970's, an association between cervical cancer and human papillomavirus (HPV) infections was postulated [3] and in the 1980's, HPV16 and HPV18 were identified from cancer biopsies of the cervix [4,5]. The HPV family contains 120 members and they are sexually transmitted. 15 viruses of these are categorized as

high risk and are considered causative agents of most cervical cancers, with over 99% of cervical lesions containing viral sequences [6].

Three HPV genes (E5, E6 and E7) are mainly participating for the origination and the progression of the cervical cancer. Their action are interrupting the negative growth regulation by host cell proteins as well as inducing instability of the genome of the infected cell [7]. The viral life cycle contains different steps including the infection of epithelial cells. HPV genomes (8 kb in length) are then established in the cells as extrachromosomal elements. The HPV genome does not encode any enzymes essential for its replication. The HPV infected cells are still replicating and in parallel to replicate its own genome, ensure the viral DNA synthesis. Capsid protein synthesis, virion assembly and release follow afterward.

The E5, E6 and E7 proteins are interacting with specific cellular factors. One of the protein family regulated by those viral proteins are the members of the retinoblastoma (Rb) family, consisting of p105 (RB), p107 and p130. The HPV E7 targets them for degradation, which result of the release and activation of E2F transcription factors that drive expression of S phase genes but also to inhibit cell growth and apoptosis mediated by p53-dependent pathway. To balance this cell growth arrest, E6 protein targets the tumor suppressor p53 for degradation. The combinatorial action of E6 and E7 proteins disrupt many cell cycle checkpoints. As a consequence, cells infected with HPV are exhibiting accumulation of cellular mutations over time and favoring progression to cancer ([8-10]) [9]. Persistent HPV infections can lead to the development of cervical cancer after several decades [11]. In many cases HPV genome are found integrated into the host chromosome. Integration of viral DNA usually silences E2 expression, which activates expression of early viral genes, including E6 and E7. The outcome of this activation is an augmentation of the proliferative capacity, which promotes the progression to cancer [12].

Figure 1: HPV oncoproteins



Molecular mechanisms by which the human papillomavirus oncoproteins cooperate to induce cervical carcinogenesis [4,5,7,13].

A crucial step toward the cell immortalization is the expression of telomerase, responsible for the DNA replication of the end of chromosomes. This activation is often present in cancers. E6 and E7 proteins contribute to this process by activating the transcription of telomerase reverse transcriptase (TERT) and inactivate RB, respectively, which together contribute to cell immortalization [14].

Viral gene expression, especially of the early genes E6 and E7, maintains tumoral cell phenotype, however they are not the only cause of this transformation. Genomic instability is one important factor. E6 and E7 proteins increase genomic instability in HPV-associated tumors, resulting in numerous chromosomal rearrangements and aneuploidy, which encompass gains or losses of whole chromosomes [15].

This genomic instability has been associated with mitotic defects, induced by the same two proteins, E6 and E7 [16]. In particular, the two early proteins induce centrosome abnormalities leading to chromosomal missegregation and the development of aneuploidy [17]. Potentially, E6 and E7 also bypass cells with abnormal mitoses to be targeted for cell death by inhibiting p53, which regulates G2-M checkpoints [18] as well as inhibition of apoptotic signal [19]. The tumorigenesis of the cells is taking many years and therefore reflecting the accumulation over time of minor chromosomal aberrations, which likely enable one cell subclone to have selective advantage over the cell population and ultimately drive the development of the tumor. Another way, leading to this observed genomic instability in HPV-infected cells, is the activation of one of the DNA damage pathway: the ATM–ATR pathway (ataxia telangiectasia- mutated–ATM and RAD3-related DNA damage repair pathway) by E6 and E7 induced-DNA damage. Indeed E6 and E7 induce DNA

damage [20] and favor the integration of exogenous DNA to the host genome [21]. This viral DNA integration could target the activation of ATM-ATR DNA damage response resulting in the amplification of integrated copies of HPV sequences. Moreover in cells containing both episomal and integrated viral genome forms, DNA re-replication triggers this DNA damage response due to the production of double and single strand breaks (DSBs and SSBs). A defective repair results in chromosomal aberrations and genomic instability [22]. Furthermore, E7 inhibits the cycle checkpoints at S or G2-M, induced by the ATM-ATR DNA damage, thus promotes mitotic entry despite DNA damage and leads to genomic instability through defective DNA repair.

Proliferation and DNA synthesis aberrations, under deprivation of growth factors trigger apoptosis mediated by P53 [23]. Rb inactivation via E7 proteins exposes the cells to p53 mediated apoptosis, and P53 itself been degraded by E6. In summary, E6 and E7 counteract the various stimuli that can induce programmed cell death. Another apoptotic pathway, associated with anchorage-independent growth and called anoikis, is also targeted by HPV proteins and allows HPV-immortalized cells to proliferate in the absence of adherence to the extracellular matrix [24].

E6 and E7 also interfere with the effects of various growth inhibitory cytokines that are induced following infection. In response to viral entry, cells produce inflammatory mediators such as tumor necrosis factor−α (TNFα).

Another layer of HPV-protein mediated deregulation targets the interferon antiviral response. E6 and E7 repress transcription of key players in the interferon pathway such as 2′-5′-oligo-adenylate synthetase 2 (OAS2) and signal transducer and activator of transcription 1 (STAT1) [25]. p53 is not only important for controlling cell cycle progression in response to genotoxic stress but also for mediating the antiproliferative effects of the IFN response. E6 overcomes the growth-suppressive effects of IFN through the formation of a complex with p53 and p300 associated with CREB-binding protein (CBP), preventing the acetylation of p53, and blocks the activation of p53-target genes, allowing continued viral replication in the presence of IFN [26]. Importantly, the binding of E6 to p300 or CBP may facilitate viral escape from immune surveillance and establish long-term persistence.

In conclusion, the occurrence of cervical cancers implies a coordinate action of HPV oncoproteins to evade normal cellular defense pathways. All these multiple-pathway targeting mechanisms are triggering cell malignancy. HPV protein actions allow the

maintenance of infected cells in proliferation in order to insure viral replication and long-term integration. In consequence, HPV-positive cells are accumulating mutations and are subject to chromosomal instability, which achieve the transformation of the infected cells to cancer cells. E6 and E7 proteins are playing a major role in this process, by disrupting p53 and Rb function as well as affecting other cellular targets such as telomerase, members of the DNA damage pathway, factors important for centrosome duplication and specific signaling proteins and in particular in immune surveillance.

1.1.3 Establishing the 1st cell line

Since the beginning of the 20th century, researchers tried to grow human cells in culture and the competition among the scientific community was extremely fervent. Animal cells were able to grow *in vitro* [27] but human cells were only proliferating for a limited amount of time.

While Henrietta Lacks was staying at the hospital, George Gey, MD, who was directing the Tissue culture laboratory and devoting his research to the establishment and maintenance of normal and patient-derived cells *in vitro* [28] and collaborators collected biopsy tissues throughout the institute. Ms Lacks' were the only cells to grow successfully in culture by using the roller-tube technique and became the first human cell line immortalized in tissue culture [29]. The cell line was named HeLa using the initial 2 letters of Henrietta Lacks' first and last names. For many years the real donor of the cell line was hidden, probably because of a missing consent form to use her cells for research purposes and the secret persisted for number of years by using names like Harriet Lane or Helen Lane [30]. The Lacks' family only discover of the existence of those cells, years after their worldwide distribution for research purposes. The evolution of the patient awareness regarding the potential use of their tissue for such applications as well as the requirement of formal patient content form is also a consequence of the HeLa history [31].

1.1.4 Dissemination and cross-contamination of cell culture

The HeLa breakthrough was considered as a major revolution in cancer research by providing unlimited cell supply. As such, the entire cancer research community coveted it and Dr Gey shipped cell samples all around the world. Many groups used this cell material to establish cell culture conditions, study how external contaminants (bacteria, virus and others) could infect human cells. The fact that the human cell culture conditions as well as the tools to discriminate the nature of two cell lines of similar morphology were not very well established allowed cross contamination of cell lines to happen in quite a number of labs. The degree of the phenomenon was revealed in 1967 by introducing genetic markers to trace cell line origins [32] and proved that HeLa was actually contaminating 18 other cell lines labeled by a different name and provided by the American Type Culture Collection (ATCC) [33]. In the 1970's, with the improvement of cell line authentication techniques and the influence of Walter Nelson-Rees, who resisted opposition from the scientific community pressures, the massive extent of the HeLa contamination across cell lines, which cast doubts upon many studies using these lines [34-36]. The problem persists and applies also to other type of cell lines. In a recent study, among 252 human cell lines, 45 (18%) of the contaminated lines were disseminated across 27 of 93 originators (29%) [37] and the problem seems to continue even today [38]. Technical developments, such as minisatellite amplification [39] and short tandem repeat profiling [40] could prevent misidentification of cell lines.

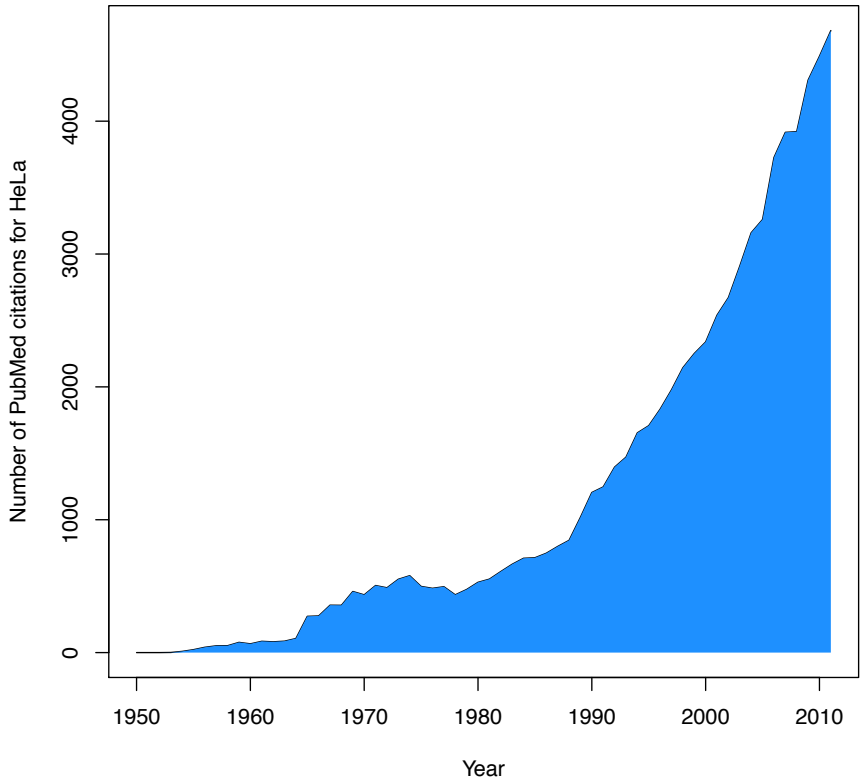
Recent examples in the literature show that this debate is still not close [8,10] and indicate that a better identification of cell line, especially of HeLa, which remains one of the major cell culture contaminant is necessary [41]. Genomic characterization of its genome could provide such markers and will potentially reduce the possibility of long-term contamination in cell culture.

1.2 60 years of research with cancer model cell lines

1.2.1 Research milestones using HeLa

Since the accessibility to cells able to grow *in vitro* in 1951 and the establishment in culture of numerous other human cancer cells, HeLa has been the most widely used human cell line in biological research. To date more than 60000 papers mention HeLa over the last 60 years of research, with roughly 4500 citations only for last year (Figure 2). The use of HeLa cells cover a wide research spectrum and its application as a model organism has contributed greatly to the characterization of important biological processes. Major milestones of HeLa use in history are presented below, before focusing on more recent works using high-throughput technologies.

Figure 2: HeLa in the literature



Number of HeLa citations retrieved from PubMed over the years.

In the 1950's and as the first human cell line, HeLa cells were used to develop standardized tissue culture good practices and rules. Cryogenics and conservation of frozen stocks were developed with HeLa. Later, cell division was closely inspected and chromosomes stained using hematocyclin to make them visible. This crucial technique allows an accurate chromosome count in human and allowed researchers to link diseases such as Down syndrome, Klinefelter syndrome, or Turner syndrome to genetic and especially chromosome number abnormality. This thus led to the development of genetic tests by amniocentesis for these particular chromosomal aberrations as well as preimplantation genetic screens applying to *in vitro* fertilization. The very rapid proliferation rate of HeLa also enabled scientists in 1954 to expand cell clones to clonal populations, which later on was derived for isolating stem cells, cloning entire animals, and *in vitro* fertilization. The same year, the company Microbiological Associates started to commercialize HeLa cells and mass-produce them.

At the beginning of the 1960's, HeLa cells traveled to space in a Soviet space capsule and NASA later also transported HeLa in their first manned mission and found that these cells grow faster in space. Five years later, the fusion of a HeLa cell with a mouse cell created the first cross-species hybrid [42]. Gene mapping thus became possible using the genetic hybrid. When the cells have fused, the nuclei fuse to form a uninucleate cell line composed of both human and mouse chromosome sets. Because the mouse and human chromosomes are recognizably different in number and shape, the two sets in the hybrid cells can be readily distinguished. However, during subsequent cell divisions, for unknown reasons the human chromosomes are gradually eliminated from the hybrid at random. The cells used are mutated for some biochemical function; therefore, if the cells are to grow, the missing function must be supplied by the other genome. At that time, HeLa applications also had a major impact on medical ethics. After the injection of HeLa cancer cells into patients in order to analyze how cancer develops, medical review boards were initiated and informed patient consent was institutionalized.

In the following decades, scientists used HeLa cells to better understand the invasiveness and infectiousness of salmonella, as well as the behavior of salmonella inside human cells [43]. Major breakthroughs, using HeLa cells, have related viral infections to disease. Nobel prizes have been awarded to Harald zur Hausen, Françoise Barré-Sinoussi and Luc Montagnier, in 2008, and to Elizabeth Blackburn,

Carol Greider, and Jack Szostak, in 2011, on research done at this period. Harald zur Hausen demonstrated that the human papilloma virus (HPV) infection played a central role in cervical cancer. Françoise Barré-Sinoussi and Luc Montagnier identified the infection mechanism of the human immunodeficiency virus (HIV). Finally Elizabeth Blackburn, Carol Greider, and Jack Szostak uncovered how telomerase prevents chromosome degradation and initiated the use of telomerase inhibitors for cancer treatment. Moreover, HeLa cells were exposed to a non-pathogenic *E. coli* strain containing *M. tuberculosis* in order to dissect the mechanisms by which the bacteria invade the cells, augment macrophage phagocytosis and survive for at least 24 hours inside the human macrophage [13], and to get insight into the infection of *M. tuberculosis* in human causing tuberculosis.Recently, the toxicity of nanoparticles [44,45] were tested on HeLa.
In summary, researchers made significant advances using HeLa cells in all of the following areas: virology, polio, scientific standards, live cell transport, genetic medicine, clones, for profit distribution of cells, space biology, genetic hybrids, ethics, salmonella, HPV, HIV, telomerase, tuberculosis, and nanotechnologies. These cells have were essential for major breakthroughs in biological and biomedical research and are still heavily depended on today.

1.2.2 The omics era

Over the last ten years, HeLa has been applied more frequently to molecular genetics studies. In particular, this cell line has been used to pioneer omics approaches such as microarrays, RNAi screens and sequencing approaches.
The evolution of DNA microarrays enabled the identification of active genes (i.e. expressed gene) at any given time or under any specific condition. The microarray is composed of a large number of gene sequences spotted on microscope glass slides. The cellular mRNA, produced during transcription, is complementary to its original portion of the DNA strand from which it was copied. The RNAs of the cells are collected, and then labeled using a reverse transcriptase enzyme, which generates a complementary DNA sequence to the mRNA. The two samples to compare are labeled using different dyes. The labeled cDNAs are hybridized to their complementary DNAs attached on the microarray slide. Each labeled cDNAs can be hybridized independently or on a common microarray. The intensity signal of each

dye is proportional the number of molecules of messenger RNA, thus, more labeled cDNAs, which hybridize to the DNA on the microarray slide. The comparison of those signals indicates in which sample the gene is more or less expressed for a specific condition.

The association of this technology and HeLa cells was used to dissect specific expression pattern involved in cell cycle, i.e. the transition between G1 and G2 phases [46]; to identify the coexpressed groups of genes involved in DNA replication, chromosome segregation, and cell adhesion [47]; to examine how histone deacetylation enzymes impact the alternative splicing of human genes [48]. Gene expression response have also been quantified in HeLa cells after exposure of the cells to stresses including heat shock, endoplasmic reticulum stress, oxidative stress, crowding, [49] and viral infection [50]. Genetic perturbations, for instance gene insertion [51] and gene overexpression or siRNA-mediated gene silencing [52], have also been assessed using microarrays.

RNA interference (RNAi) has become a widely used method for studying gene function in higher eukaryotes by silencing gene expression using synthetic antisense oligonucleotides. It exploits a physiological mechanism that represses gene expression, primarily by causing the degradation of mRNA transcripts. In mammals, RNA is processed into shorter fragments (siRNAs) by dicer enzymes. These siRNAs are unwound and one strand of the siRNA duplex, called the guide strand, is integrated within the multi-protein RNA-induced silencing complex (RISC). When the guide strand binds to a complementary mRNA sequence and triggers its degradation by Argonaute. Argonaute contains the catalytic activity of the RISC complex, which results in a post-transcriptional gene silencing.

RNAi mechanism has been exploited as a research tool, both in cell culture and in living organisms, by selectively and robustly inducing suppression of specific genes of interest. It therefore can be used to identify genes involved in biological pathways. An example of its application is the characterization of receptor-associated protein 80 (RAP80) as a BRCA1-interacting protein in humans, by recruiting BRCA1 to DNA damage sites and which functions in G2/M checkpoint control [53]. RNA interference screens in HeLa have led to the functional classification of pathways involved in cell division [54], mitosis/cytokinesis [55], and endocytosis [56].

RNAi can also be used for high-throughput screening (HTS) that systematically shuts down each gene in the cell, which can help identify the genes associated with specific

biological phenotypes. One of the major advantages of the genome-scale RNAi screening is its ability to simultaneously interrogate thousands of genes. This approach have yielded catalogs of genes that mediate cell division [57] as well as an array of basic cellular phenotypes ranging from apoptosis to proliferation [58].

The human genome sequence was sequenced by the automated Sanger method [59]. This method is considered a 'first-generation' technology and more recent methods are referred to as next-generation sequencing (NGS). The new generation of methods parallelizes the sequencing process, producing massive amount of sequences simultaneously for low costs. Different strategies are on the market, which use a combination of template preparation, sequencing and imaging [60]. The template preparation implies common steps between the different technologies, including randomly genomic DNA fragmentation into smaller sizes from which either fragment templates or mate-pair templates are created. The two template libraries differ in the length of the DNA fragmented during the sequencing process (<1 kb for fragment templates and >2 kb for mate-pair library - Figure 6). The template is usually immobilized to a solid surface to be subsequently sequenced. The sequencing methods also differ between platforms. The application of NGS covers many different areas in biology. These include variant discovery, de novo assemblies, analysis of the cell transcriptome, tissues and organisms (RNA-seq) [61], genome-wide profiling of epigenetic marks and chromatin structure using other sequencing-based methods (ChIP-seq, methylseq and DNase-seq) [62], and metagenomics studies [63]. The transcriptome of HeLa has also been characterized with next generation sequencing technologies (e.g. poly(A) RNA [64] and small RNAs [65]), and analysis by deep RNA sequencing has enabled the first comparison of the proteome and transcriptome for a single cell type [66]. These genome-wide technologies have changed the type of biological question that we can ask and they became central in research nowadays.

While such studies have led to breakthroughs in modern molecular biology, they were designed and analyzed in the absence of genomic sequence information for the HeLa cell line. Rather, researchers have been forced to use the human reference genome, despite its clear differences from that of a cancer cell line that has been evolving in the laboratory for several decades. Indeed, compared to the human reference genome, substantial chromosomal aberrations in the HeLa cell line have been revealed by cytogenetic methods [67-76]. A combination of these techniques (comparative genomic hybridization (CGH), fluorescence in situ hybridization

(FISH) and spectral karyotyping (SKY)) has been used to determine the karyotype of a CCL2 HeLa cell line [75]. This cell line contained 2 subclonal populations, which were both hypertriploid (3n+) and contained a variable total number of chromosomes (76-80) and a variable number of abnormal chromosomes (22-25) per cell. The comparison of their spectral karyotype with previously published G-banding karyotypes [67-72,74,76] and FISH study [73] indicates high concordance between independent measurements of chromosomal aberrations in HeLa. These well-documented genomic aberrations underscore the need for a HeLa reference in the age of genomics.

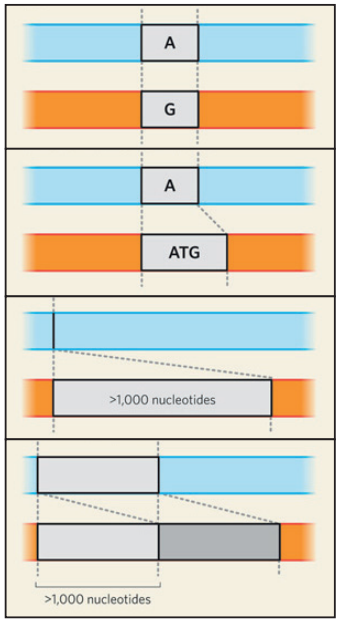
1.3 Genomic variations

Human genetic variations can be classified according to their respective base pair length impacted (Figure 3). Structural variations (SVs) were originally described as insertions, deletions and inversions larger than 1 kb in size [77,78]. In addition, with the recent development of sequencing technologies, SVs also include much smaller events (>50 bp) representing deletions, novel insertions, inversions, mobile-element transpositions, duplications and translocations as well as unbalanced structural variations, which are defined as variations changing the number of base pairs, resulting from an abnormal number of copies of one or more fragments of the DNA (copy number variations - CNVs), [79,80]. Smaller events (<50bp) are considered as short insertion-deletion variants (indels) and single-nucleotide variations (SNVs).

The discovery and genotyping of the full spectrum of genomic variations has been extensively developed in the past years, with the use of sequencing technologies notably, and is currently central to understanding their impact on human disease, complex traits and evolution.

Figure 3: Genomic variant classes

The different types of genomic variants are classified as single-nucleotide variation (SNV); insertion or deletion of several nucleotides (indel); insertion or deletion of thousands of nucleotides (SV); and duplication or multiplication of DNA segments more than 1,000 nucleotides long (CNV). Adapted from [81]



1.3.1 Structural variants

SVs play a role in evolution and, currently, it is demonstrated that two human individual differ by 5000-10,000 inherited SVs [82,83]. Acquisition of specific SVs leads to disease phenotypes. In cancer for example, most of the tumoral tissues present numerous genomic rearrangements [84-88]. Among SVs, CNVs remain challenging to call, despite the identification of different CNVs across individuals [80,89,90]. Those difficulties arise due to the multiplicity of the approaches used that can greatly modify the outcome [91]. Numerous CNV genotypes have been associated with complex traits, such as a 20 kb deletion upstream of the IRGM gene with Crohn's disease [92], a 45 kb deletion upstream of NEGR1 with body mass index [93], a 32 kb deletion that removes two late-cornified envelope genes with psoriasis [94], and a 117 kb deletion of UGT2B17 with osteoporosis [95]. The mechanisms causing CNV incorporate meiotic recombination, homology-directed and non-homologous repair of double-strand breaks, and errors in replication [96].

Microarrays have been a method of choice to identify CNVs. One of the methods is the array comparative genomic hybridization (array CGH), which is based on competitively hybridizing fluorescently labeled test and reference samples to a known DNA sequence spotted on a solid glass support and then analyzing the

hybridization ratio [79]. Another technology used is the SNP microarrays, in which the target DNA sequences are discriminated on the basis of a single base difference. Each sample is hybridized on an independent array, followed by SNP genotyping and copy-number analysis [97]. These hybridization-based microarray approaches are limited in sensitivity as they cannot provide information on the location of the duplicated copies and it is very difficult to map nucleotide-resolution breakpoints. Both array CGH and SNP platforms assume a diploid reference genome, which is not verified in the presence of duplicated sequences. Moreover the diversity of array platforms as well as the different analysis pipelines confer significant result differences, depending on the methods used [98-101].

With the recent development of sequencing, NGS is commonly used in order to detect such SVs. The main advantage of NGS is the capacity to capture all possible SV classes with a single sequencing experiment. The detection methods can be summarized in four different approaches (read-pairs, read-depth, split-read and sequence assembly). They all first map of the sequenced reads to the human reference genome and second identify specific read signatures representative of different classes of SVs (Figure 4). Different approaches are used to identify SVs with NGS data, nevertheless each method is not comprehensive and a combination of different tools is required.

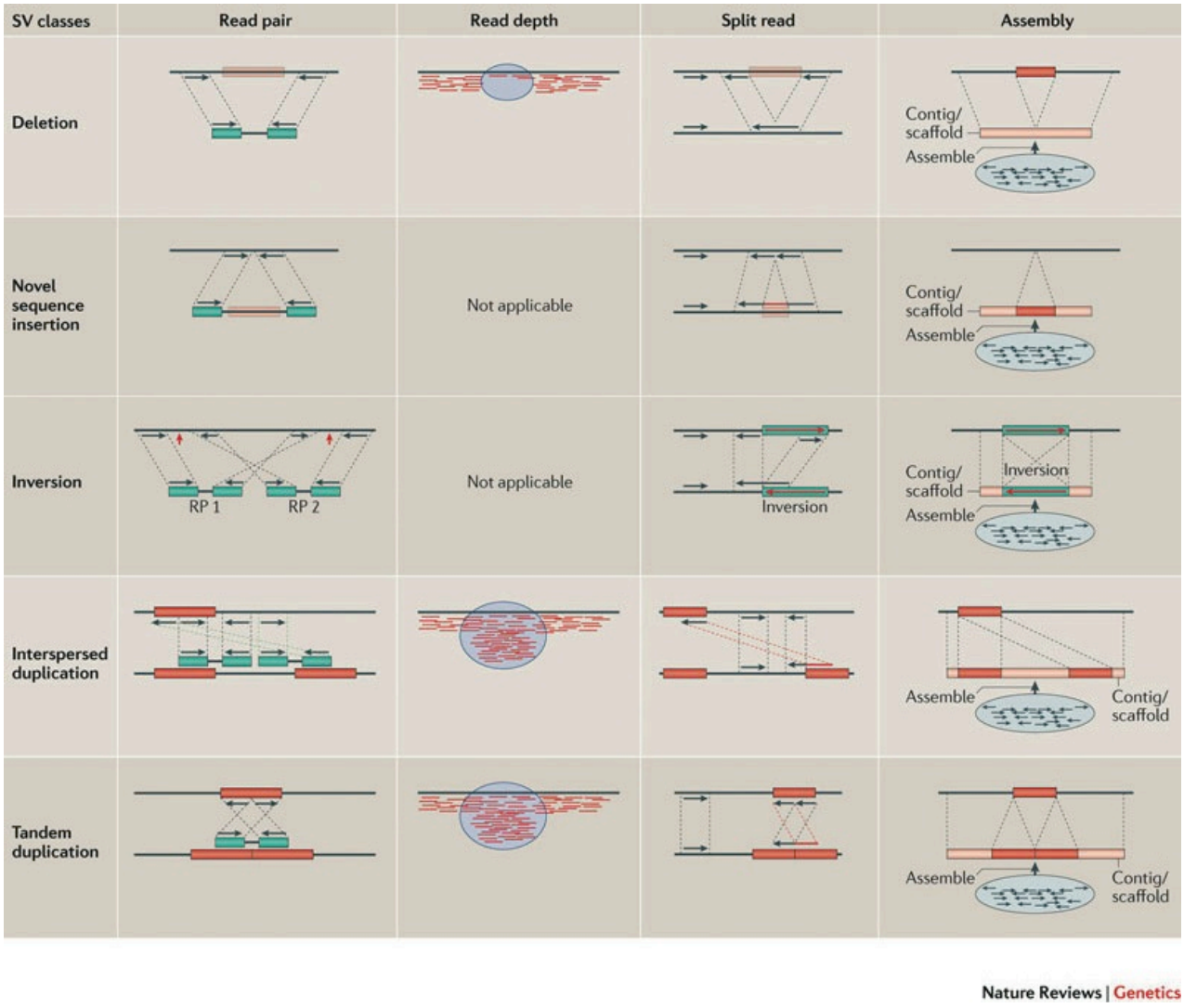
When paired-end reads are generated by sequencing from the start and end of the same DNA molecule, read-pair methods use the span and orientation of the reads to identify aberrant pairs in which the distance between the reads and/or the orientation of the read pairs are discordant with the human reference genome. When aligning the reads on the human reference genome, pairs of read mapping too far apart, suggest deletions, whereas if the distance between the two reads is too close, insertions can be suspected. Read pair orientation inconsistencies identify inversions and tandem duplications.

Read-depth analysis uses coverage values for each genomic position. The global idea is that duplicated regions will show significantly higher read depth and deletions will show reduced read depth when compared to diploid regions. Using multiple human individual genomes, the absolute copy number has been estimated using this approach [83].

Detection of exact breakpoints of all SV classes is possible using split-read algorithm by scanning the reads alignment on the reference genome and identifying any break in the alignment and report a 'split' sequence-read signature [102].

Finally, assembly methods can potentially detect all classes of SVs with base-pair resolution breakpoints, but the assembly of short read sequences is not trivial and can result in complex contig fragmentation (especially in repetitive regions), which is very difficult to solve [103].

Figure 4: Structural variant read signatures

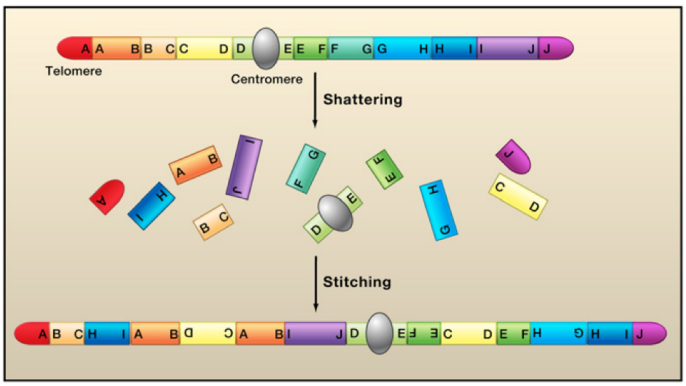


Read signatures used in SV detection (adapted from [104])

1.3.2 Chromothripsis

The possibility to sequence deeper and cheaper enables the discovery of a new type of complex genomic rearrangements such as chromothripsis. This phenomenon is characterized by numerous genomic rearrangements involving several chromosomes with alternation between a few CN states [85]. The mechanism generating such events is hypothesized to comprise double-strand breaks followed by reassembly through end-joining in a random manner (Figure 5). The occurrence of these multiple double-strand breaks is still under investigation and a possible explanation proposes a defect in mitotic chromosome segregation, mediated by the formation of micronuclei, resulting in chromosome pulverization [105]. Loss of DNA material could occur during the formation of this complete reshuffled chromosome segments. Chromothripsis seems to occur in 2–3% of all cancers [85]. Oncogene amplification could be also associated with rearranged segments and confer a selective advantage to the affected cancer cell.

Figure 5: Chromothripsis



A complex chromosomal rearrangement formed by chromothripsis, with shattered segments followed by stitching back process. Chromosomal segments are indicated as blocks with different letters and shading, and the orientation by an arrow in each block. Adapted from [106].

1.3.3 Single Nucleotide variants

In human, two individuals differ in one base every kilobase [107,108]. These single-nucleotide variations, depending on their location, could potentially impact the function or structure of the proteins, which could result in inherited monogenic disorders. However, most SNVs are located in non-coding regions of the genome,

and have no direct known impact on the phenotype of an individual. These SNVs are useful as markers in population genetics and evolutionary studies [109].

During the last decade genome-wide association studies (GWAS) have identified hundreds of common SNVs (with a frequency of ~5% or higher in the general population) associated with a couple of hundred traits or diseases. Nevertheless, most of the complex traits that have been studied using this approach have reasonable levels of heritability, but the proportion of that heritability explained by very common variants is surprisingly low [110]. Possible reasons have been proposed. One view considers that the genetic variation contributing to common traits has subtle effects, is context-dependent and specific combinations of common variants might have substantially greater effects on risks than individual common variants. Rarer variants (frequency of 1-5% in the general population) have also supporting evidence to play a role for certain diseases, including a prostate-cancer-associated SNV that is upstream of a gene and is associated with expression level in cell lines [111], and breast-cancer-associated intronic SNVs, associated with expression and binding of transcription factors [112].

The constitution of catalogs of very common variants in the population and less common variants contributing to common diseases has successfully been established [113]. If much of the causality is due to rare variants, such an approach is unlikely to uncover much more of the genetic control of diseases than has already been revealed by the current GWA efforts. The sequencing of entire genomes or entire exomes could solve the issue and has already begun to be used at a small scale. Such studies will eventually be carried out, depending on the affordability, similarly to GWA studies, with very large sample sizes, providing sufficient statistical evidence to implicate variants on the basis of association evidence alone. Applications of this concept have already been undertaken. The sequencing of the exomes of four unrelated cases with Freeman–Sheldon syndrome and eight controls showed that the causal variants contained at least one coding indel or non-synonymous or splice-site variant in all four cases that was not present in any of the controls, nor in SNV repositories [114]. Another example applied whole-exome sequencing of a patient with an undiagnosed disorder, to identify a few homozygous missense mutations sitting in a gene known to be associated with a congenital chloride-losing diarrhea, which correspond to the symptoms [115].

Among the different improvements needed to make this approach feasible, methods for SNV detection will need to progress, as currently there are tens of thousands of false-positive variant calls per genome [116], which subsequently confuses the determination of the causal variants. In addition, one key remaining challenge is the association between SNVs and functional consequences on the impacted gene. The SNV functional effect prediction methods are based on the assumption that disease-causing mutations are more likely to occur at positions that are conserved throughout evolution, as compared with positions that are not conserved. This assumption suggests that prediction could be based on sequence homology or disease-causing amino-acid substitution. Therefore they have common structural features that distinguished them from neutral substitutions, indicating that structure could also be used for prediction. Based on these two approaches, different prediction software packages have become available to identify disease causal variants [117].

2 Aim of the project

The HeLa cell line is the very first and most used human cell line in current biological research and its application as a model organism generated major contributions to the characterization of important biological processes. Indeed, since the 1950's, all type of studies have played a significant role in modern molecular biology. Yet their design and analysis have been carried out in the absence of sequence information for this cell line. Rather, researchers have been forced to use the human reference genome, despite the clear differences from a cancer cell line that has been evolving in the laboratory for several decades. In the light of this widespread application of HeLa, the absence of sequence information for this cell line is problematic, since it constrains scientists to use the human reference genome and transcriptome to design their experiments (probes) and analyze their data (alignments). Given the important use of this cell line in genetic, molecular or genomic studies, it is crucial to take into account the differences between the human reference genome and the HeLa genome concerning the design and the interpretation of any experiments using HeLa.

The aim of my PhD project is to characterize in detail the genomic and transcriptomic landscape of a HeLa cell line, in order to provide the community with a comprehensive resource for this cell line. We used next generation sequencing to generate high-coverage DNA dataset and report the copy number (CN) for each position of the genome, as well as single nucleotide variants (SNVs) and structural variants (SVs). We also profile the HeLa transcriptome, and assess expression differences between HeLa and normal human tissues, by comparison to the Illumina Human Body Map 2.0. We explore the impact of CN heterogeneity on gene expression and show a significant shift in expression level profiles depending and the CN. We also investigate the cell cycle pathway and the DNA repair pathways in more detail, integrating different layers of information. The analysis suggests that key regulators involved in the checkpoint of the cell cycle are affected by mutations. Meanwhile the cells sense the unstable genome and activate the DNA repair pathways but they remain ineffective due to defective components. Taken together, these results constitute the first view of the genomic landscape of the most widely used cell line in molecular biology research. Our characterization of the HeLa genome and transcriptome yield a set of resources that will inform the design of future experiments and the re-interpretation of old studies.

3 Methods

All resources mentioned in the subsequent sections are available at the following URL, where accession numbers are updated: http://steinmetzlab.embl.de/helaseq/
In order to login, please use the following information:
Username: helaseq / Password: helaseq

3.1 Cell line, culture conditions and DNA/RNA preparation

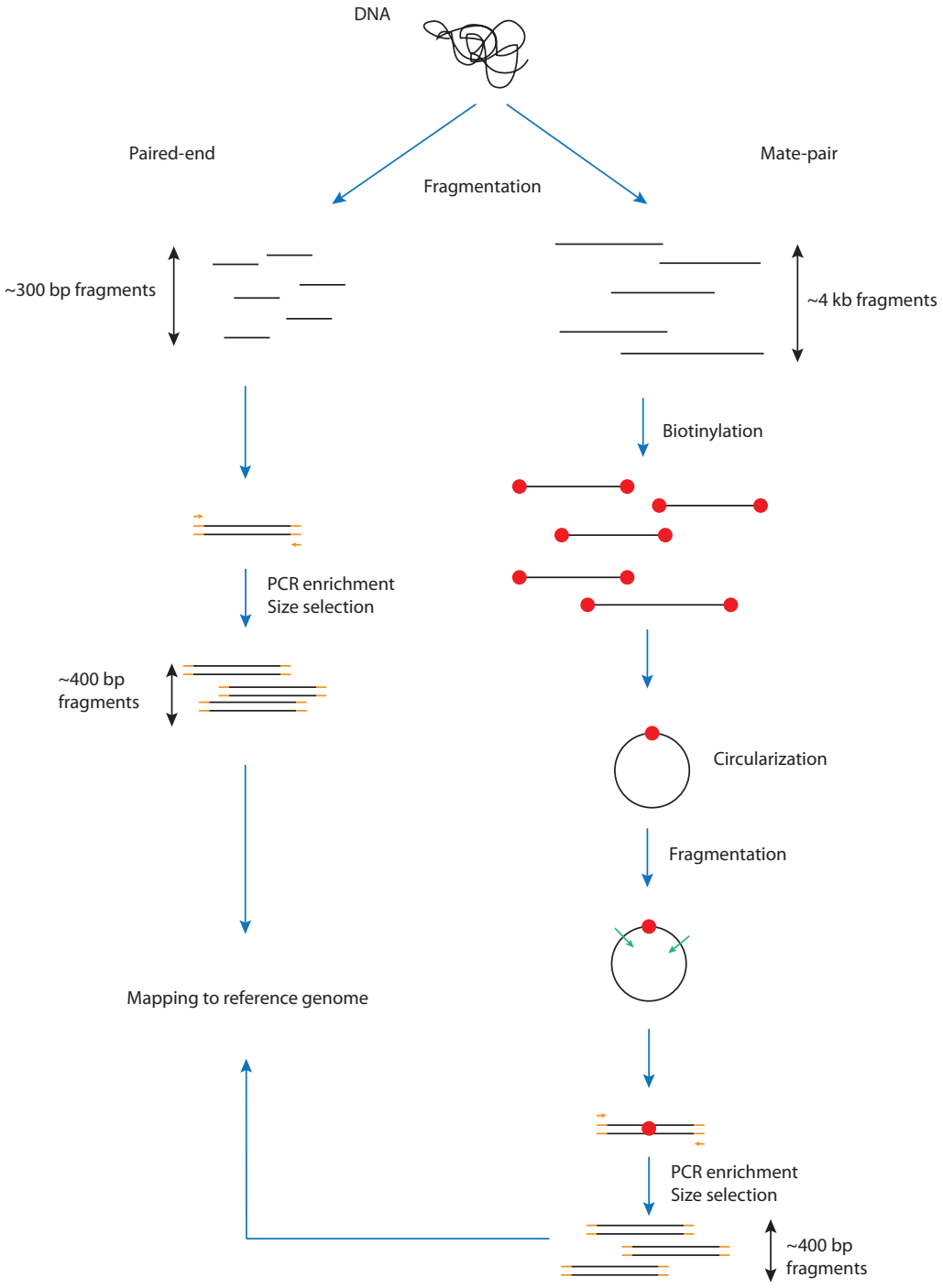
The cell line, HeLa H2B-mRFP and mEGFP-α-tubulin, was derived from HeLa Kyoto background [118]. The cell can be purchased in CLS Cell Lines Service GmbH (Catalog number 300670). It has been used to identify genes associated with cell division [57]. Cells were cultured in DMEM (Gibco) containing 4.5 g/l glucose (Sigma), 10 % heatinactivated foetal bovine serum (Sigma), 2 mM glutamine (Sigma), 100 U/ml penicillin and 100μg/ml streptomycin (Sigma), and incubated at 37°C and 5% CO2. Cells were harvested at 80% confluency. RNA-free genomic DNA was prepared using Qiagen DNeasy Blood and Tissue kit (Qiagen). Total RNA was extracted using Trizol.
The cell line was tested negatively for mycoplasma contamination using the MycoAlert® mycoplasma detection kit (Lonza).

3.2 Library preparations

3.2.1 DNAseq library preparation

RNA-free genomic DNA was prepared using Qiagen DNeasy Blood and Tissue kit (Qiagen). For short insert size paired-end libraries (PE), the genomic DNA was sheared using Covaris S2. The sequencing library was prepared following the manufacturer's protocol (Illumina) [119] using NEBNext DNA Sample Prep Master Mix Set 1 (NEB).
For long insert size mate-pair libraries (MP), 10ug of genomic DNA were sheared using the Hydroshear (GeneMachines) and libraries were prepared using the Illumina MP v2 reagents and protocol.
The paired-end and the mate pair sequencing approaches are illustrated in Figure 6.

Figure 6: Paired-end and mate pair sequencing



For paired-end sequencing, DNA is fragmented into smaller pieces (200 to 500 pb), adapters are linked to the extremities of the DNA fragment and the product is sequenced from both ends. After alignment of the generated reads against a reference genome, the mapped reads are pointing inwards.

In the mate pair approach, the DNA is fragmented in larger fragment (2 to 5 kb). Biotins are added to both end of the fragment, which would be circularized. A fragmentation of these ligated fragments lead to the generation of smaller pieces (400 to 600 pb) containing the biotin label. Adapters are then linked to the extremities of those fragments and the sequencing is performed the same way as the paired-end method. After mapping, the reads are pointing outwards.

3.2.2 Strand-specific RNA-seq library preparation

Poly-A RNA isolation, the RNA fragmentation and cDNA synthesis protocols are described by Yoon *et al.* [120]. The cDNA was processed for DNA library prep according to Parkhomchuk *et al.* [121]. In summary, the protocol includes RNA fragmentation, first strand synthesis followed by second strand synthesis using dNTPs and dUTPs. End repair, A-tailing and ligation are then achieved as well as size selection for a specific fragment size (300 to 350 bp). The second strand is cleaved by hydrolysis of uracil in the dsDNA. The resulting strand specific cDNA is amplified before sequencing.

3.3 Sequencing and alignment

3.3.1 DNAseq

The PE library was sequenced on eight lanes of HiSeq 2000 (Illumina) using the manufacturer's recommended pipeline (v1.1). The raw reads were submitted to the European Nucleotide Archive (ENA) and the accession number is available from our website[3]. The resulting 101bp paired-end reads (1.1 billion) were mapped with GSNAP [122] to the human reference genome (GRch37). Only unique alignment for each read was reported (-n1 -Q), otherwise default parameters were used. 86 % of the total read number (946 millions) was aligned.

The MP library was sequenced on one lane of HiSeq 2000 (Illumina) as above. 81% of the total number (383 millions) was aligned, resulting in a 155x physical coverage after redundancy filtering. The MP raw read sequences are also available from our website[3].

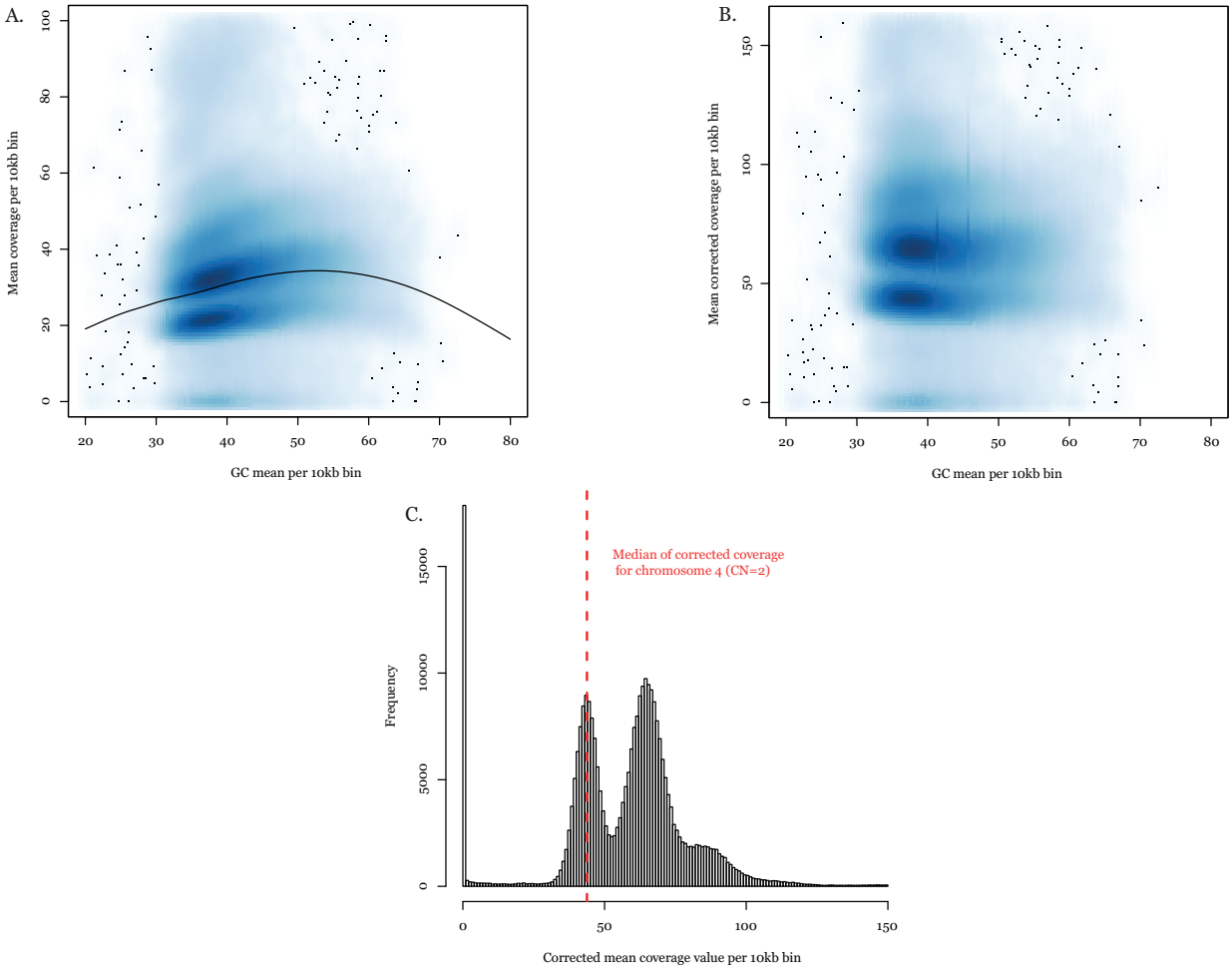
3.3.2 RNAseq

The three libraries were sequenced on nine lanes of Genome Analyzer II (Illumina) using the manufacturer's recommended pipeline (Illumina - v1.18). Two paired-end read lengths were generated: 76bp and 105pb. The raw reads were deposited in Sequence Read Archive (SRA) and the accession number is available from our website[3]. RNA sequencing reads (450 millions) were mapped with GSNAP [122] to the HeLa genome. Accession numbers are available from our website[3]. Like the DNA alignment, we report only unique alignment (-Q -n 1) otherwise default parameters were used. 56 % of the total read number (253 millions) was aligned.
Raw reads from the BodyMap 2.0, provided by Illumina (ENA number: ERP000546) were downloaded and aligned with GSNAP using the same parameters as above.

3.4 DNAseq analysis

From the fully preprocessed alignment (bam file) we computed the depth of coverage of the HeLa genome in 10 kb bins. To this we applied a mappability correction by dividing the count for each bin by the proportion of mappable positions in that bin. If this proportion was below 0.5 we discarded the counts for this bin and assigned it the value 'NA'. A position in the genome was called mappable if a simulated read of length 101bp (the length we used in our DNAseq experiment) starting at that position had exactly one valid alignment when processed with our alignment pipeline as described in Section 4.3.
We also corrected for the GC-content dependent coverage bias. We used a local fit to describe the relation of GC-count to coverage per bin and the ratio between that fit and the desired coverage of 60 bp as the correction factor. Figure 7 shows the effect of this correction.
To describe the extent of CN aberrations of HeLa we created a track segmenting the genome according to integer CN, which was obtained with the R/Bioconductor package DNACopy [123] followed by mixture model fitting.
Since DNACopy works with log2-ratios we transformed our corrected depth-of-coverage data into a log2-ratio by dividing by 30 (an arbitrary choice) and taking the log2. We then normalized by subtraction of the median log2-ratio of a manually curated region of CN 2 from the HeLa genome (Chromosome 4).

Figure 7: GC correction effect on the coverage



A. The panel describes the mean coverage value per 10 kb bin (y-axis) as a function of the GC mean value per 10 kb bin, before any correction was applied. The black line represent a robust local regression fitted on the data.
B. The panel describes the corrected mean coverage value per 10 kb bin (y-axis) as a function of the GC mean value per 10kb bin.
C. Distribution of corrected mean coverage value per 10 kb bins along the genome. The dashed red line corresponds to the corrected coverage median value for the chromosome 4 used to normalize against the corrected mean coverage value for regions of CN 2.

3.4.1 Copy number segmentation with DNAcopy

We used the standard workflow of DNACopy to generate a segmentation of our log2-ratio data using the "sdundo" option to require at least 2-standard deviation difference between the medians of adjacent segments.
From the segmented log2 ratio s we calculated the copy number track $c = 2 \cdot 2^s$ to which we fitted a mixture model of m normal distributions:

$$X \sim \sum_{i=1}^{m} w_i \mathcal{N}(i, \sigma_i)$$

Each of those normal distributions corresponded to a CN state and has the CN as its mean.
The following parameters define our mixture model:
m – the number of components
σ_i – the standard deviation of the i-th mixture component
w_i – the weight of the i-th mixture component

We then assigned each segment an integer copy number based on its probabilities in the normal distributions given by our mixture model. To do so we calculate for each segment s the log-odds ratio for each possible CN state.
Given $\varphi_{\mu,\sigma}(x)$ as the probability density function of a normal distribution with mean μ and standard deviation σ and k as the integer CN state, the probability of the segment in each of the normal distributions was defined as follows:

$$\Pr(s|k) = \varphi_{k,\sigma_k}(s)$$

The probability of being in state k given value s is defined as (Bayes-Theorem):

$$\Pr(k|s) = \frac{\Pr(s|k) \cdot \Pr(k)}{\Pr(s)}$$

Here $\Pr(s)$ can be omitted since it is a scaling factor only and using the weights w of the mixture model fit as prior-probability we get $\Pr(k|s) = w_k \cdot \varphi_{k,\sigma_k}(s)$.

The confidence of the highest probability call was assessed by its log odds ratio with the sum of probabilities in all other CN states, which we used as an estimate of the probability of s not to be in state k:

$$\Pr(\neg k|s) = -\Pr(\mathrm{k}|\mathrm{s}) + \sum_{i=1}^{m} \Pr(i|s)$$

$$\mathrm{logodds}(\mathrm{k}|\mathrm{s}) = \log_{10}\left(\frac{\Pr(k|s)}{\Pr(\neg k|s)}\right)$$

We accepted the highest probability CN as the state of segment s if the log odds ratio was higher than 1.27, which corresponds to a confidence in the call of at least 0.95; if the log odds were lower than this cut-off we assigned NA as the CN state.

3.4.2 Single nucleotide variants and small indels

In order to call SNVs, we used the pipeline described in the best practice variant detection with the GATK v3[2], which consists in duplicate removal, base quality score recalibration and indel realignment. We used a minimum confidence score threshold of 30 as filtering parameter for the GATK UnifiedGenotyper tool [124].
Short indels (1 to 50 bp) were called using the program PINDEL [102] with default parameters and insert size of 302 bp, estimated from a sample of a million read pairs.
The loss of heterozygosity (LOH) track is based on the distribution of the allele frequencies of all SNVs called by GATK. To identify homozygous regions in the HeLa genome we calculate the proportion of homozygous SNV calls in 100kb bins and call a bin homozygous if the proportion of homozygous calls is > 0.5. To identify big blocks of homozygous regions we apply a segmentation algorithm to the binned proportions of homozygous calls (DNAcopy) and then classify the segments based on the same criterion as above (homozygous proportion > 0.5) to obtain a track that segments the genome into homozygous and heterozygous blocks. This approach was applied in the same way for our SNV calls as well as for the SNV calls performed by the HapMap consortium [125] on 3 individuals: NA12878, NA12891 and NA12891.

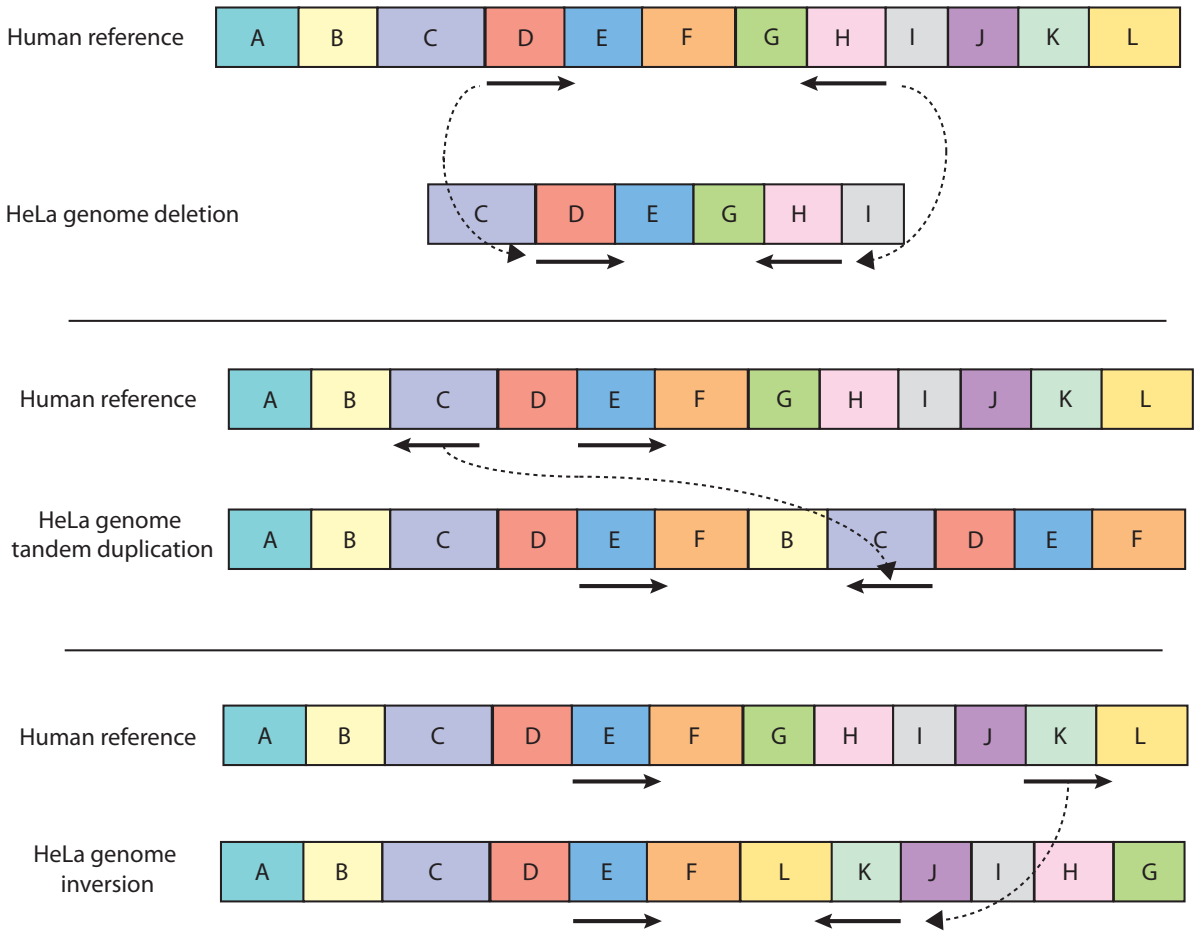
3.4.3 Large structural variants

Structural rearrangements were detected using paired-end mapping [126,127]. Paired-end methods rely on two key characteristics of a sequenced paired-end or mate-pair library: (1) the expected mapping distribution and (2) the default orientation of the read pairs with respect to each other.

The structural rearrangement calls were filtered using phase I 1000 Genomes Project (http://1000genomes.org) genome data as well as germline data of 30 additional whole-genome sequencing samples [128] to exclude polymorphic germline structural variants as well as rearrangement calls caused by mapping artifacts. We only considered for further analysis those rearrangements, which were present in at most 0.5% of the 1000 Genomes Project samples assessed and not in the additional germline samples. Two rearrangement calls were considered to be identical, hence constituting a likely germline variant if they displayed an overlap in terms of genomic coordinates with their end coordinates differing by less than 5 kb. Furthermore, rearrangement calls with less than 3 supporting pairs as well as supporting pairs with average mapping quality less than 20 were excluded for further analysis.

Three inversions, 93 deletions, 52 tandem duplications and 12 translocations were randomly picked in order to be validated by polymerase chain reaction (PCR). PCR primers were designed for predicted SVs with Primer3 (parameters: Tm=60°C; Tmin=57 °C; Tmax=63 °C; optimum length=25bp; minlength=18bp; max-length=26bp; mingc=40; maxgc=60). Primers matching to repeat database [129] were excluded. The primers designed for deletion, inversion and tandem duplication events are following the rules described on Figure 8. The primers validating translocations are spanning the breakpoint-junction-sequences of predicted SVs. 10 ng each of HeLa DNA were amplified in 30 µl PCR reactions using 0.3 µl of Phire Polymerase (Thermo Scientific - F-122S), 5X Phire reaction buffer, 200 µM dNTPs and 0.5µM of primers for 36 cycles. PCR cycle includes initial denaturation at 98°C for 30s, denaturation at 98°C for 10s, annealing at 60°C for 10s, extension at 72°C for 2min and final extension for 5min at 72°C. The products were run on 0.8-1% agarose for determining the size.

Figure 8: Primer design



For deletions, the amplicons generated are smaller than expected size. The orientation of the primer targeting tandem duplication and inversion allow detection of an amplicon only if the event is present.

3.4.4 Multicolor fluorescent *in situ* hybridization (M-FISH)

Metaphase spreads of HeLa cells were performed as follows. Subconfluent HeLa cells were arrested in metaphase by adding colcemid (final concentration 0.270 μg/ml) to the culture medium for 2.5 h at 37° C in 5% CO_2. Cells were washed with PBS, treated with trypsin-ethylenediaminetetraacetic acid (EDTA), resuspended into cell medium and centrifuged for 5 min at 1200 rpm. To the cell pellet in 3 ml of PBS, 7 ml of a prewarmed hypotonic 0.0375 M KCl solution was added. Cells were incubated for 20 min at 37° C. Cells were centrifuged for 5 min at 1200 rpm and the pellet was resuspended in 3–5 ml of 0.0375 M KCl solution. The cells were fixed with methanol/acetic acid (3:1, vol:vol) by gently pipetting. Fixation was performed four

times prior to spreading the cells on glass slides. For M-FISH analysis, chromosome–specific painting probes were labeled using DOP-PCR amplified DNA using 7 different fluorochromes in a combinatorial manner and hybridized as previously described [130]. Twelve metaphase spreads were acquired by using a Leica DM RXA RF8 epifluorescence microscope (Leica Mikrosysteme GmbH, Bensheim, Germany) equipped with a Sensys CCD camera (Photometrics, Tucson, AZ). Camera and microscope were controlled by the Leica MCK-FISH software (Leica Microsystems Imaging solutions, Cambridge, United Kingdom). Metaphase spreads were processed on the basis of the Leica MCK software and presented as multicolor karyotypes.

3.4.5 Virus integration detection

We aligned the DNAseq reads obtained from HeLa to a genome consisting of the human reference (GRCh37) and a set of known whole virus genomes obtained from the viral genome resource (NCBI) [131]. Potential virus insertions are indicated by read pairs with one read mapping to a chromosome from the human reference and one read mapping to a virus genome. We extracted all those read pairs and performed a clustering based on overlap to find clusters of read pairs indicating virus insertions. To account for mapping artifacts introduced by similarities between virus genomes we also clustered read-pairs together where the reads on the human genome overlap and the reads on the virus genomes map to viruses from the same family (papillomavirus, herpesvirus, adeno-associated virus, adenovirus, lentivirus, poxvirus and retrovirus).

3.5 RNAseq analysis

3.5.1 Expression level and copy number

The number of reads per gene were counted using HTSeq [Anders, http://www-huber.embl.de/users/anders/HTSeq/]. The per gene counts were divided by the length of the gene and then we transformed these normalized counts per kb of genes. The log_{10} of the count per kb of genes is considered as a proxy of expression.
We called SNV on the RNA sequencing data using the same pipeline used for the DNA sequencing data following the best practice variant detection with GATK v3[2].

Allelic SNV counts in genomic regions of CN 3, were extracted from the output of the GATK caller.

3.5.2 Comparing HeLa to the Illumina Body Map 2.0

The BodyMap 2.0 data (ENA number: ERP000546), generated on HiSeq 2000 instruments (Illumina), consist of 16 human tissue types, including adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells. The 16-tissue raw reads were downloaded and aligned by GSNAP with the parameters that were used for the HeLa transcriptome sequencing reads. To explore which functions are specific to HeLa compare to other human tissues, the number of reads per gene were counted like described in the previous section for both samples and compared using the DESeq package [132]. We used DESeq for normalization and preprocessing (variance stabilization, size factor correction) and estimated the physiological range of counts for each gene from the 16 body map samples. We obtained z-scores (z) for our HeLa samples compared to the distribution described by the 16 Body Map samples. With the estimated standard deviation $\bar{\sigma}$ and mean $\bar{\mu}$ of the 16 Body map samples and x being the vector of means of the three normalized HeLa RNASeq counts, the z-scores were computed as follows:

$$\mathrm{z} = \frac{x - \bar{\mu}}{\bar{\sigma}}$$

We used a cut-off of 3 to determine which genes were significantly overexpressed compared to the physiological range. We qualified non-expressed genes those with a mean of less than 1 count.

We search for enriched terms using MGSA [133] with 10 independent runs of the Markov chain of 10^8 steps each. For every parameter, we used a regularly spaced grid with 11 points. Default search intervals for the model parameters turned out to be not appropriate because the maximum of the posterior was often reached at the bounds, implying that the most likely fits lied outside the search intervals. Thus, the search intervals for the parameters p, alpha, and beta were set to [0.001, 0.01], [0.001, 0.05] and [0.7, 0.9] respectively for the highly expressed genes, and [0.0001, 0.02], [0.001, 0.2], and [0.7, 0.95] for the non-expressed genes.

4 Results

All resources mentioned in the subsequent sections are available at the following URL, where accession numbers are updated: http://steinmetzlab.embl.de/helaseq/. In order to login, please use the following information:
Username: helaseq / Password: helaseq

4.1 Genomic Landscape

Our analysis reveals the extent and nature of the differences between the human reference genome and the HeLa genome. We identified genetic changes including SNVs and SVs in the HeLa Kyoto genome by DNA shotgun sequencing and by comparing the resulting 1.1 billion sequence reads of length 101 nt to the human reference genome (GRch37). A genome-wide representation of our results is presented in Figure 9.

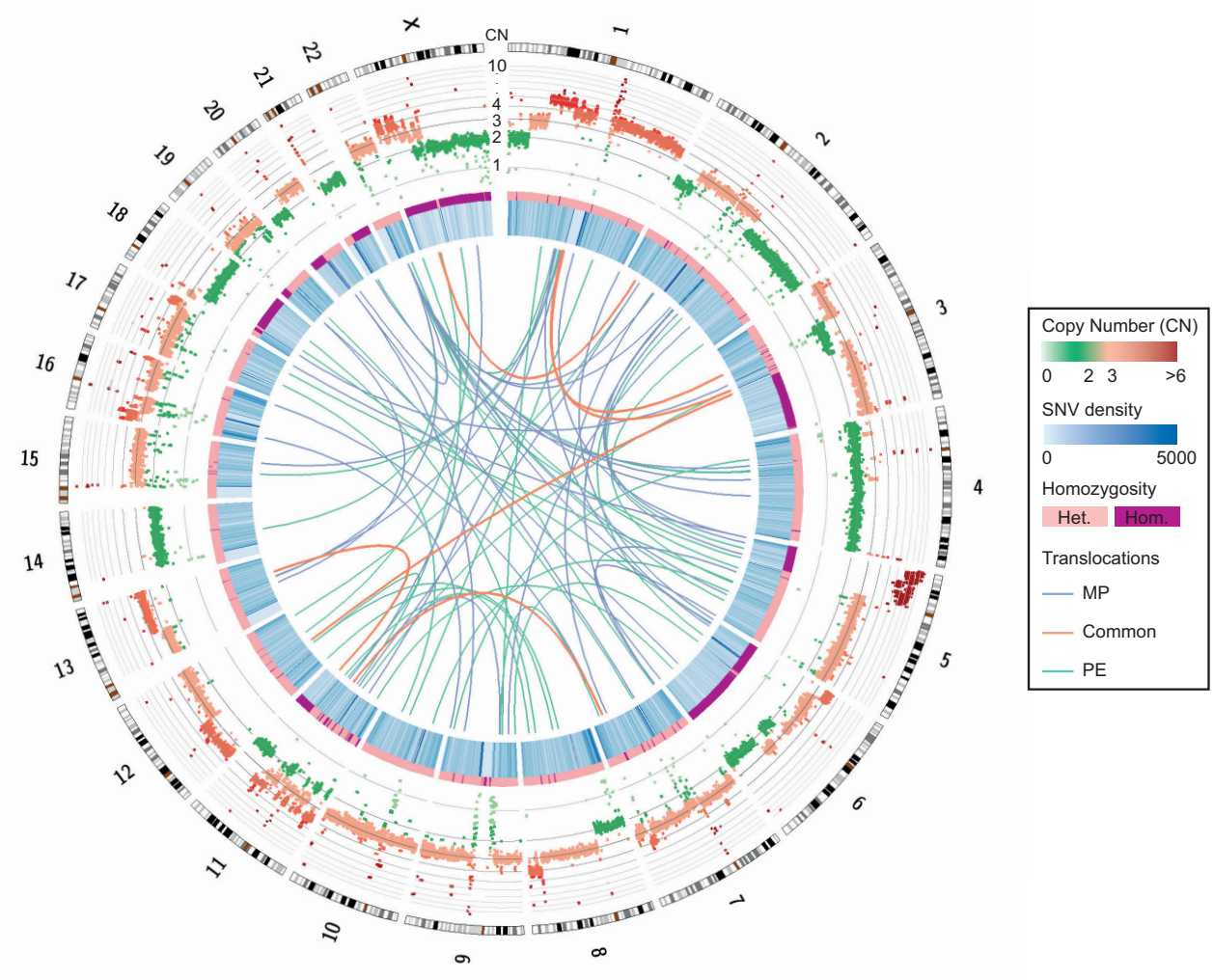
Numerous CN changes, as well as sequence modifications, were observed at the single nucleotide level and in larger structural rearrangements; these variations are detailed in the following sections. We report a compendium of genomic variation (CN, SNVs and SVs) as well as the first HeLa genome draft is available as VCF and FASTA files respectively, available on our website[3]. By integrating a set of high-confidence homozygous variant calls (SNVs, small indels as well as large deletions and insertions – Table 1) into the human reference (GRch37) we constructed a HeLa genome sequence, which is available on our website[3]. We retained the overall chromosome structure and encoded copy number aberrations in a separate file. We provide a tool to perform the translation of coordinates between GRch37 and our HeLa reference, and report our variant calls in both coordinate systems, also available from our website[3].

Table 1: Homozygous variants

Classes	Homozygous calls used to build the HeLa genome
SNVs	1733577
Large Deletions (DELLY)	748
Short Deletions (PINDEL)	14964
Short Insertions (PINDEL)	3446

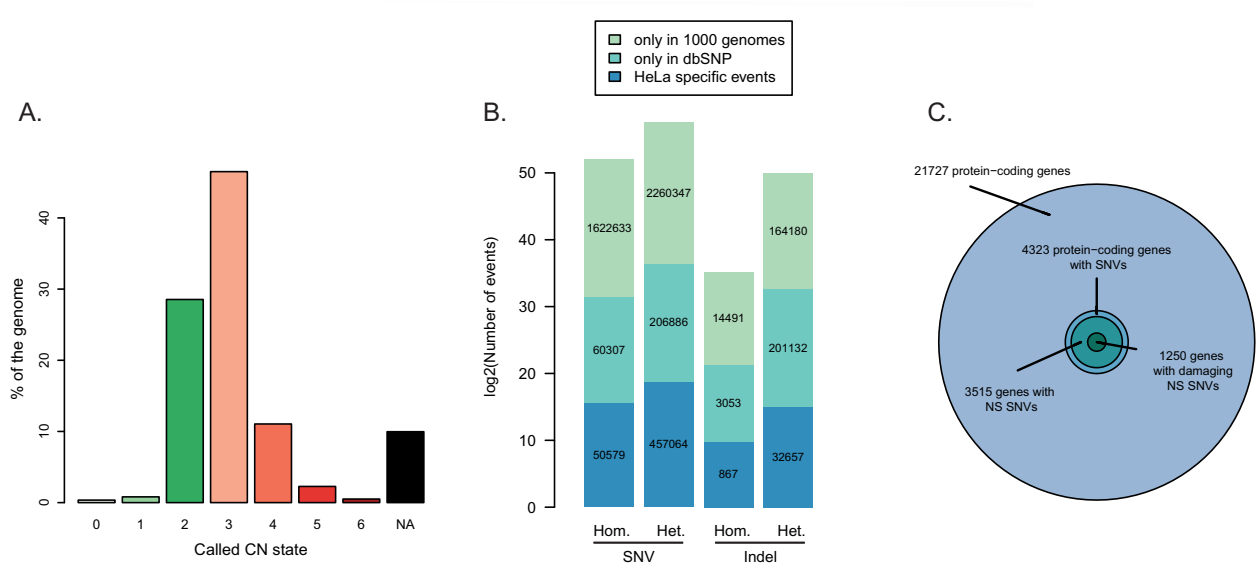
Summary table of homozygous calls (SNVs, large deletions, short deletions and short insertions) integrated in the Human reference genome to build the HeLa genome. The methods used are mentioned in brackets.

Figure 9: Overview of the HeLa genome



Each of the chromosomes (1-22 and X) is depicted along the outside of the circle as ideograms. Data tracks are depicted in concentric rings. The outermost track contains read-depth plots (10 kb-binned coverage) in colored dots. The color scheme is light green for CN1, green for CN2, and red colors, starting for light to dark are representing CN values from 3 to 10. The CN values (1 to 10) are depicted in circular grey lines. The homozygosity track indicates potential homozygous regions (darker purple) based on the allelic distribution of our SNV calls. The SNV density track indicates the distribution of SNV across the genome (darker blue for higher density). Connection lines are showing interchromosomal translocations called with paired end sequencing data (dark orange), called with mate-pair data (dark yellow) and called by both dataset (purple). We used Circos to generate this plot [134].

Figure 10: CN, SNV and functional impact



A. CN distribution across the genome.
The histogram depicts the proportion (in percentage) of genomic position associated with each CN value observed in HeLa. CN value 0 corresponds to positions in the genome with coverage less than half of what we expect for CN 1 (reasons can be either poor mappability or deletion events). A CN value of "NA" indicates a segment where the CN could not be determined with confidence ≥ 0.95 (see Methods Section).
B. SNV and indel classification.
Overlap of our SNV calls and indel calls with dbSNP, 1000 genome project. The y-axis shows the log2-transformed counts. The four different classes of events represented on the x-axis are homozygous ("Hom.") and heterozygous ("Het.") SNVs (called by GATK) and homologous ("Hom.") and heterozygous ("Het.") short indels (called by PINDEL)
C. Protein-coding gene classification.
The subgroups are protein-coding genes containing SNVs, non-synonymous SNVs, damaging non-synonymous mutations.

4.1.1 Copy number by read depth analysis

By inferring CN using sequencing read depth, we found extensive CN heterogeneity across the HeLa genome, with most loci present in three copies (Figure 9 and Figure 10-A). We adjusted the observed coverage levels for systematic effects of GC-content and mappability. We then applied a segmentation algorithm coupled with a gaussian mixture model to assign each segment to its most probable integer CN state, resulting in an assignment of each region of the human reference genome to a discrete CN (Figure 10-A). These results corroborate previous observations that the genome displays an unbalanced number of chromosomes (3n+) [75], while providing a high-resolution view of genome-wide CN state. A complex pattern of genomic rearrangements was expected from previously conducted karyotyping by low resolution FISH.

Interestingly, some genomic regions display frequent CN alternation, which is indicative of chromothripsis, a recently discovered phenomenon observed in some cancers that has never been reported for HeLa cells [135-137]. These results are described in further detail in the following sections.

4.1.2 Single nucleotide variants and indels

We next catalogued the SNVs in HeLa and used them to infer allelic variability and potential functional consequences of the variation in this genome. SNVs between the HeLa and reference genomes were called by applying the GATK [124] standard preprocessing steps (duplicate removal, base quality score recalibration and indel (insertion/deletion) realignment), followed by genotyping with the GATK UnifiedGenotyper. Short indels (1 to 50 bp) were called using the program PINDEL [102]; larger indels were classified as structural variants (SVs).

We found 1,733,577 SNVs and 18,411 indels that were homozygous in HeLa. 97.1% and 95.3% were already reported in dbSNP respectively (release 137). Among these homozygous calls classified in dbSNP, 96.4% of SNVs and 82.6% of indels were also contained in the 1000 genomes dataset (Figure 10-B). These variants thus represent common variants in the human population. The remaining 53,121 variants are either specific to Henrietta Lacks, somatic mutations of the tumor, or arose during transformation and propagation of the cell line. The lack of samples from the donor and tumor precludes us from distinguishing these possibilities. Among the SNVs classified in dbSNP, 1,121,149 overlap with Venter [138] and Watson [139] SNV lists (928,157 and 855,005 respectively).

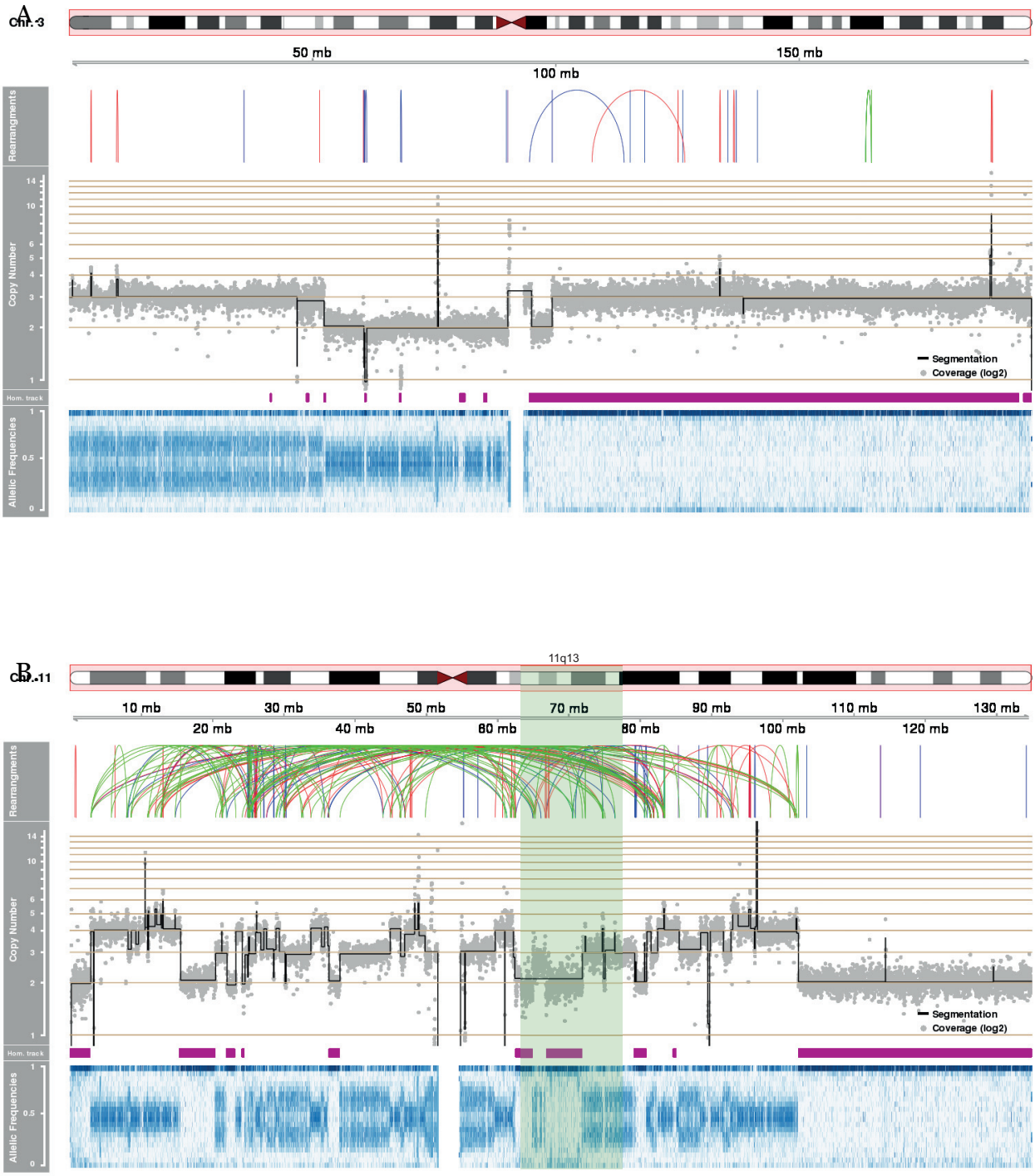
In addition, extensive allelic variability exists in HeLa. We detected 2,924,297 heterozygous SNVs and 397,969 heterozygous indels. 84.4% and 91.8% were respectively already reported in dbSNP. Among these, 91.6% heterozygous SNVs and 44.9% heterozygous indels were contained in the 1000 genomes dataset, and thus represent common variants ([140]-B).

We used SIFT [141] to predict the effect of our called SNVs on protein function. Among the 4,657,874 SNVs, 29,881 are localised in coding regions and within this subset 4865 constitute nonsynonymous mutations, 1442 of those are predicted to alter protein functionality (Figure 10-C). These potentially damaging SNVs are overlapping 1250 genes. The GO class “sensory perception of chemical stimulus” is enriched in this subset of genes. This could be the result of the deregulation of pathways that rely on external

chemical stimuli by constitutively activating enzymes downstream of the receptor, which is a common mechanism of cancer [140]. The list of mutations having an effect on protein function is available from our website[3].

Using the allele frequencies of our SNV calls, we created a classifier to identify homozygous regions of the genome (100 kb bins). The regions called homozygous indicate potential loss of heterozygosity (LOH). The validation would require comparison to the unavailable germline cells of Henrietta Lacks, which would filter out the constitutional genome variation. We analyzed the distribution of minor and major allele frequency of our SNV calls together with the average copy number. Overall, in HeLa, 23% of the genome is exhibiting LOH and therefore ¼ of the genome carry only one gene version, which could be still functional or impacted by disrupting event. The exact same approach was performed on genotyped individuals from the HapMap project [125] and could not reveal any LOH regions in their genome. Chromosome 3, which contains a large LOH region, is represented in Figure 11-A (all chromosome plots are available in the Appendix A). It is interesting to note that many of the homozygous segments in the HeLa genome correspond to previously reported LOH cervical cancer hotspots, namely on chromosomes 3p, 6p, 11q and 18q [142-147]. This suggests that these LOH events arose during the cervical cancer, prior to cultivation of the HeLa cell line.

Figure 11: SVs, CN, LOH for chromosome 3 and 11



The top panels (Rearrangements) contain lines representing the predicted connections between fragments derived from SV calls based on read pair orientation and spacing. Different read pair signatures indicate the following event types: deletions (blue), tandem duplications (red), inversions (green), or interchromosomal translocations (purple). The center panel (Copy Number) represents the 10 kb binned coverage (grey) overlaid by the segmentation for each bin (black) normalised to CN=2 in logarithmic scale. The associated CN is shown in the y-axis. Potential LOH events (Homozygosity) are indicated in dark purple.
The bottom panel (Allele Frequencies) depicts a heatmap of the allele frequencies for each 10 kb bins.
The region 11q13 is marked in shaded green.

4.1.3 SVs

In order to maximize the detection of large SVs (>50 bp) in the HeLa genome, it was necessary to integrate paired-end and mate pair sequencing data. Our SV calls were made with the method DELLY [126] as well as PINDEL (for inversions and tandem duplications). The two sequencing technologies (paired-end and mate pairs) produce pairs of nucleotide sequences from both ends of a DNA fragment. The size of this fragment can vary depending on the technique used: paired-end reads effectively pinpoint short (around 300 bp) SVs at high resolution, while mate pair outperforms paired-end technology for identifying larger (2-10 kb) SVs.

We obtained single-nucleotide resolution breakpoints for 2279 (paired-end) and 614 (mate pair) of the SVs implicated by the merged clusters. A summary of all the called SVs using paired-end and mate pair data can be found in the Table 2.

Our SV analysis reveals several highly interconnected regions of the HeLa genome coinciding with pronounced CN alternations; these features together strongly indicate chromothripsis [136].

Table 2: Overview of the SVs

Classes	DELLY calls using PE data		DELLY calls using MP data		PINDEL calls using PE data	
	number of calls	overlap (%)	number of calls	overlap (%)	number of calls	overlap (%)
Deletions	1882	14.88	234	67.09	-	-
Duplications	312	35.90	191	70.16	4488	19.52
Inversions	33	72.73	139	77.70	7140	13.03
Translocations	102	5.88	50	12.00	-	-

Number of deletions, duplications, inversions and translocations called using two methods (DELLY and PINDEL) and two different sequencing data, paired-end (PE) and mate-pair (MP). For each category, the overlap is calculated as numbers of calls overlapping at least one of two other categories.

Among the subset of deletions (93), tandem duplications (52), inversions (3) and translocations (12), 35%, 50%, 100% and 80% of the events are respectively validated by PCR. Multiplex fluorescent *in situ* hybridization (M-FISH) is a molecular cytogenetic technique allowing the simultaneous visualization of all human chromosomes in different colors (Figure 12). Chromosome-specific probe pools (chromosome painting probes) are generated by degenerate oligonucleotide-primed polymerase chain reaction and incorporating fluorescently labeled nucleotide. The use of combinatorial labeling

scheme permits to distinguish clearly chromosomes after *in situ* hybridization. Due to the resolution of the technique, only large (5 to 10Mb) chromosomal rearrangements can be detected. The chromosomal painting of 12 cells reveals common rearrangements but also events that are only occurring at the single cell level. In addition, the number of chromosomes by cells is on average is 64 with a minimum of 62 and a maximum of 68. In addition, 20 large interchromosmal translocations occurred on all of the 12 cells and 11 of them are also present in the translocation calls from DNA sequencing data. These results emphases the number of large DNA rearrangements during the cell culture. It is important to note that this chromosomal snapshot of DNA content does not imply the fixation in the population of such DNA reshuffling. However those events could potentially provide a selective advantage specific for specific subclones. If this can be true, it is probably very rare. Indeed our genomic analysis of CN do not indicate that the population of cells sequenced was heteregenous as we could assign for most of the genome a specific CN state and we do not observe any coverage value in between two CN states for large DNA segments.

Figure 12: Colored Karyotype by M-FISH

Multicolor FISH karyotype of 12 HeLa cells in metaphase uncovers common events, which are found in the 12 cells analyzed (top panel) and single cell events (bottom panel).

4.1.4 Viral insertions

Cervical cancer is often associated with insertion of human papillomavirus and especially HPV16 and HPV18. We assessed if the integration was still present in the genome of cell line. We found an insertion of HPV18 on chromosome 8 (Appendix A), which corroborates with previous characterization of the integration site of HPV18 [75].
In addition we report insertions of 9 additional integration sites (Appendix B). The viral sequence similarity between the members of a specific family is adding difficulty to disentangle which virus is actually incorporated in the sequence. Among the 7 families (papillomavirus, herpesvirus, adeno-associated virus, adenovirus, lentivirus, poxvirus and retrovirus) investigated, only 3 viral types are reported: herpesvirus, poxvirus and papillomavirus. One of the potential integration sites for the human herpesvirus 5 on chromosome 12 disrupts the gene TUBA1C, a component of the protofilaments of the microtubule.

4.1.5 Chromothripsis

The massive rearrangements observed in some parts of the HeLa genome are a hallmark of chromothripsis, a phenomenon observed in cancer cells, where parts of chromosomes are shattered and re-arranged at random. Chromothripsis has been associated with 2 to 3% of all cancers [135]. The signature of chromothripsis includes: (i) high number of CN switches along a chromosome, (ii) alternation between 2 or 3 different CN states (for instance on chromosome 11, the CN varies between 2 and 3), (iii) high interconnectivity among the fragments (i.e., regions which are usually far apart on a chromosome are connected), and (iv) double-minute chromosomes, indicated by small genomic regions of the same very high CN connected by aberrant read pairs [135-137]. Although we found no evidence for the presence of double-minute chromosomes in HeLa, chromosomes 5, 19, X and especially 11 clearly display properties (i)-(iii), indicating at least one chromothripsis event in the HeLa genome (Figure 11-B).
The complex intrachromosomal rearrangements on chromosome 11 have previously been observed at low resolution using cytogenetic analysis (DAPI and G-banding) [75]. Furthermore, chromosome 11 presents indications of LOH according to our allelic distribution analysis (Figure 11-B). In previous studies where LOH on chromosome 11

was also observed, the introduction of a functional copy of chromosome 11 into HeLa suppressed its characteristic aggressive proliferation phenotype. This indicates the presence of tumor suppressor genes on chromosome 11, and has also been observed in other cervical cancer cell lines [74,148-152]. A potential cervical cancer-suppressor gene has been mapped to the 11q13 region [153], which in HeLa displays rearrangements symptomatic of chromothripsis (Figure 11-B). Deletions and LOH within this region have also been associated with neuroblastoma as well as breast, head and neck, and nasopharynx cancers [154-159]. It is therefore possible that chromothripsis and LOH on chromosome 11 contributed to the development of Henrietta Lacks' cervical cancer.

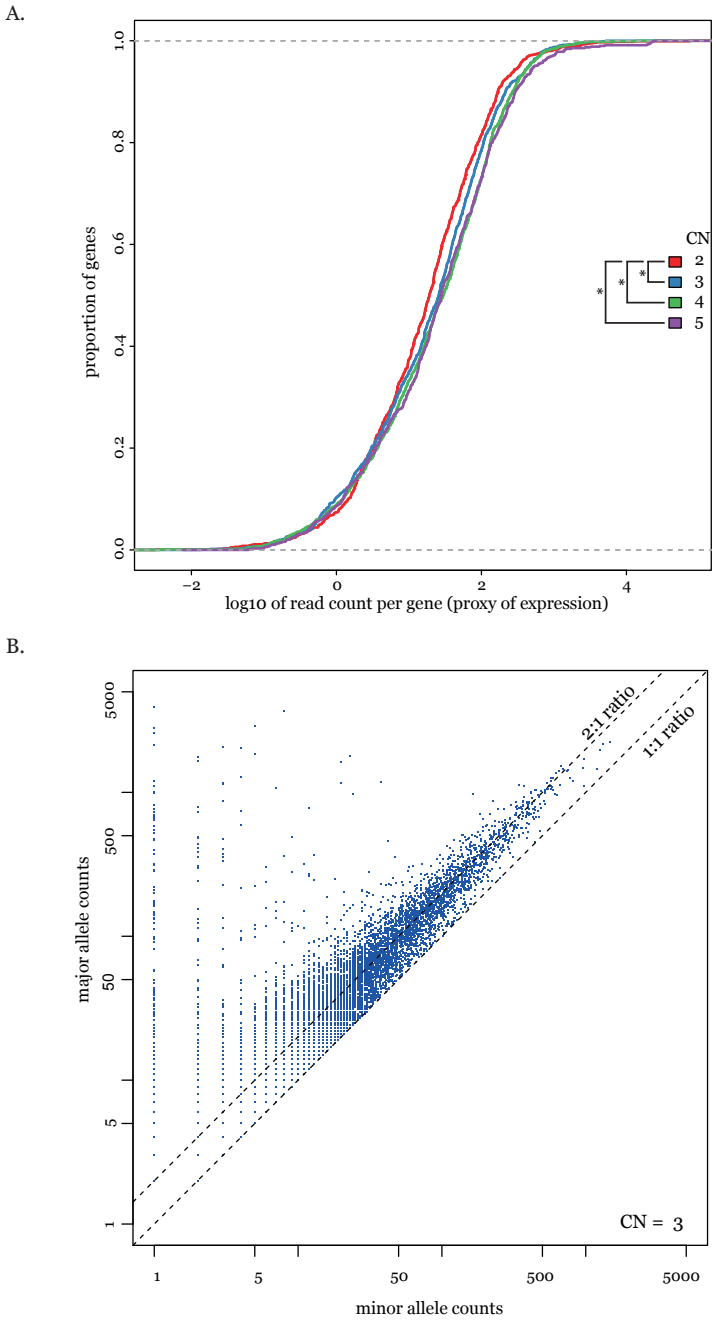
4.2 Transcriptomic Landscape

In order to analyze in more details the expression of such genome, we sequenced polyA RNAs from HeLa cells which produced nearly 400 millions of 76 nt and 101 nt paired-end reads. The raw reads are available in the ENA database under the accession number ERP001421. 63% of the total read number (250 millions) was aligned to the HeLa reference genome that we created and described previously. This resource is accessible in order to retrieve information on the expression of particular genes.

4.2.1 Expression level and copy number

How do HeLa cells deal with high CN? We see significant overall increase in expression levels as a function of gene copy number (Figure 13-A; Wilcoxon test p-value < 0.01). In the range of highly expressed genes (log10 read counts per kb of gene >=2), we can see a shift in the CN state for 3, 4 and 5 compared to CN 2. Furthermore, we studied regions with a CN of 3 and with a major allele frequency of 2, and measured the allele frequency in the corresponding transcripts. If dosage compensation was achieved by shutting off one of the three copies of each gene, allelic ratios of 1:1 or 2:0 would be expected for such genes. However, we found that the genomic ratio of 2:1 is maintained in the RNA expression data (Figure 13-B). These data suggest that there is no gene dosage compensation in HeLa, but if it occurs, then it would have to influence all alleles equally and not very strongly. Figure 13-A, however, suggests that this is not the general trend. In order to determine if particular gene expression levels are affected by CN and sequence variation in HeLa, it is necessary to compare these levels to a physiological reference.

Figure 13: Gene expression stratified by CN and allelic expression in CN3



A- The plot shows the ecdf (empirical cumulative distribution functions) of the expression value for all genes present for each CN. The x-value show the log10 read counts per kb of gene and the y-axis shows the corresponding cumulative density, which corresponds to the percentage of values smaller or equal to the given x-value. The shift right along the x-axis of all CN states greater than 2 indicates that expression levels increase with CN overall. We perform Wilcoxon test to compare the distribution of expression between CN states and * indicates significant difference (p-value < 0.01).

B- For each SNVs localised in CN3 genome segments, we plotted the allele having the higher number of counts (major allele on the y-axis) as a function of the allele having the lower count (minor allele on the x-axis). The two dashed lines represent allelic count ratio 2:1 and 1:1.

4.2.2 Comparing HeLa to the Illumina Body Map 2.0

To gain a better understanding of the adaptation of the gene expression profile of HeLa through years of selective pressure towards fast proliferation, we compared our HeLa transcription data to publicly available RNASeq data from the Illumina Human Body Map 2.0 project (ENA number: ERP000546). We compared the gene expression levels (calculated as per gene counts) of the 16 tissues from the Body Map with those obtained from our three HeLa RNAseq libraries.

We found 1907 genes that were more highly expressed in HeLa than in any tissue represented by the body map, and 19579 genes that were not expressed in HeLa. To functionally characterize the 2 sets of genes, we looked for over-represented Gene Ontology (GO) categories [160] using MGSA [133], a method which substantially limits the number of redundant enriched categories. The large number of genes that is not expressed by HeLa is consistent with the view that these are no longer utilized by HeLa or have been silenced by mutations that were not selected against. Among the GO terms enriched in the non-expressed genes (Table 3), we find terms related to response to stimulus (defense response, immune system response, G-protein coupled receptor signaling pathway), communication within a cell, or between cells (ion transport), or within a multicellular organism (multicellular organismal process) and protein cleavage (proteolysis). The genes that are more highly expressed are consistent with functions beneficial to the cell line, such as proliferation (cell cycle phase), transcription (RNA processing, rRNA transcription) and DNA repair (Table 4).

Functions that are not expressed in HeLa are thus likely poor candidates for further study in this cell line. The transcriptome data reveals further insights that will be useful in design and interpretation of functional studies using HeLa.

4.2.3 Cell cycle and DNA repair pathways analysis

In the pool of genes highly expressed compare to normal physiological of 16 normal human tissues, an enrichment of genes playing a role in DNA repair mechanism is observed. *A priori* this statement is contradictory with regard to the sequence variation identified in HeLa especially SVs and CN variations. We integrated sequence variation information and expression value to describe in finer resolution what is happening for

individual player of this pathways. Figure 14 corresponds to the KEGG representation [161] of the cell cycle pathway.

Table 3: GO term enrichment for non-expressed genes in HeLa.

GO id	GO term (Biological Processes)	Total number of genes	Number of genes in lower expressed group	Posterior Probablity	Standard deviation
GO:0006952	defense response	1127	418	0.60	0.16
GO:0006508	proteolysis	964	269	0.60	0.16
GO:0007186	G-protein coupled receptor signaling pathway	1148	736	0.60	0.16
GO:0006811	ion transport	985	379	0.60	0.16
GO:0032501	multicellular organismal process	5708	2070	0.60	0.16
GO:0002376	immune system process	1710	524	0.54	0.15

Table 4: GO term enrichment for highly expressed genes in HeLa.

GO id	GO term (Biological Processes)	Total number of genes	Number of genes in highly expressed group	Posterior Probablity	Standard deviation
GO:0006281	DNA repair	382	75	0.94	0.00
GO:0006396	RNA processing	667	89	0.69	0.01
GO:0022403	cell cycle phase	807	163	0.68	0.02
GO:0009303	rRNA transcription	19	5	0.52	0.01

Globally, the players involved in the cell cycle network are most of the time expressed in high level compared to the physiological range of expression (when comparing to the BodyMap data). In addition, more than a dozen genes contain damaging NS mutations and are non-expressed in HeLa. These two indications likely reflect an inactivation of those players. Most of the genes in this subset are acting as negative regulators for different cell cycle phase entry such as mitosis (WEE [162]), S phase (E2F ORC, MCM and CDC7 [163]), G1 phase (P19 [164]). More global regulators are also part of this group like P53, P300, GADD45, TGFβ [165] and 14.3.3 protein complexes [166]). All together, these data suggest a potentially uncontrolled activation of the cell cycle leading to constant proliferation.

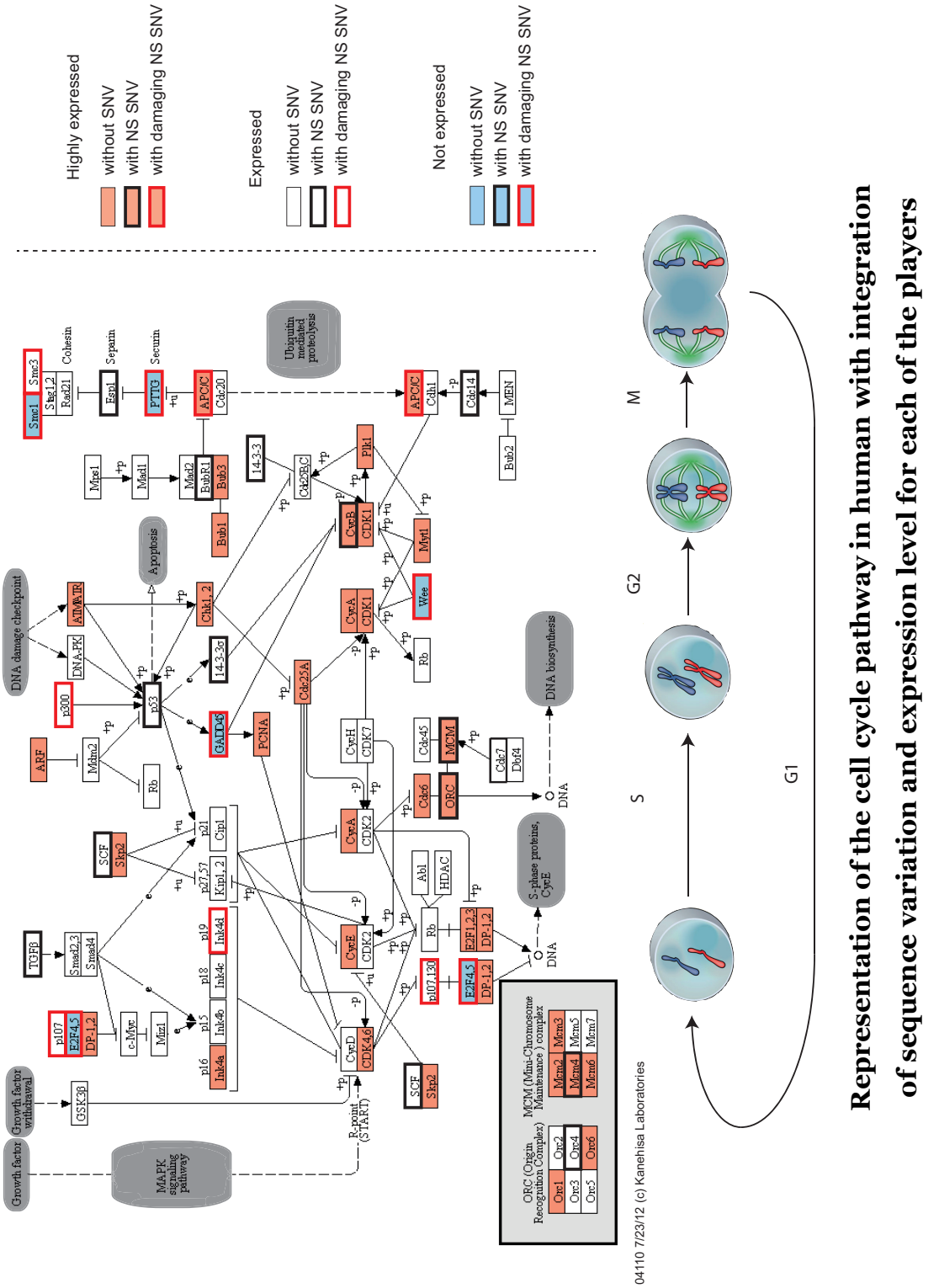
Normally P53 integrates inputs, which evaluate genomic damage and stress conditions and subsequently trigger apoptosis if the conditions are not optimal. As observed in HeLa, a weaker P53 activity, due to an identify NS mutation, leads to inappropriate replication of cells and consequently allows cells containing genomic damages or inappropriate DNA content to replicate. In addition of this P53 alteration, the genes

involved in DNA repair pathways are highly expressed compared to the physiological range of expression (Table 4), however each of those pathways contain at least one predicted non-functional gene (Figure 15 and Figure 16). These observations suggest that HeLa is still able to sense the damage occurring on its DNA and try to fix it by activating all dedicated pathways but they remain inefficient due the some damaged components.

P53 mutations are reported to happen in almost every type of cancer at different frequency (from 6% in cervix to 42% for specific ovarian cancer - IARC P53; release R15; [167]) and it is likely that this mutation and perhaps others in this pathway can be attributed to the tumor development *in vivo*.

Every repair mechanism is specialized for particular DNA damages [168]. Those pathways can be categorized in two groups depending on the type of DNA lesions that they are fixing. The first group, which contains homologous recombination (Figure 15-A) and non-homologous end-joining (Figure 15-B) is specialized to fix toxic double-strand DNA breaks (DSBs) [169] and the second group, formed by nucleotide excision repair (Figure 16-A), base-excision repair (Figure 16-B) and mismatch repair (Figure 16-C), tends to repair base impairing or DNA single-strand break (SSBs). Despite the fact that most of the players of the different pathways are expressed compared to the physiological expression range, we notice that at least one player for each the DNA repair mechanism is disrupted by a potential damaging NS mutations. This suggests that the cells sense the unstable genome and activate the pathway but the pathway remains ineffective due to defective components.

Figure 14: Cell cycle



These three figures present the KEGG representation for the cell cycle pathway (Figure 14), the double strand break repair pathways (Figure 15) and the single strand break pathways (Figure 16). The background color of each of the genes (represented as boxes) is color-coded depending of it is highly expressed (orange), normally expressed (white) or not expressed (blue) compare to its expression level across 16 different normal human tissues. The border of the box is color-coded depending if the gene is predicted to contain NS SNV (thick black) or NS damaging SNV (thick red).

Figure 15: Double strand breaks repair pathways

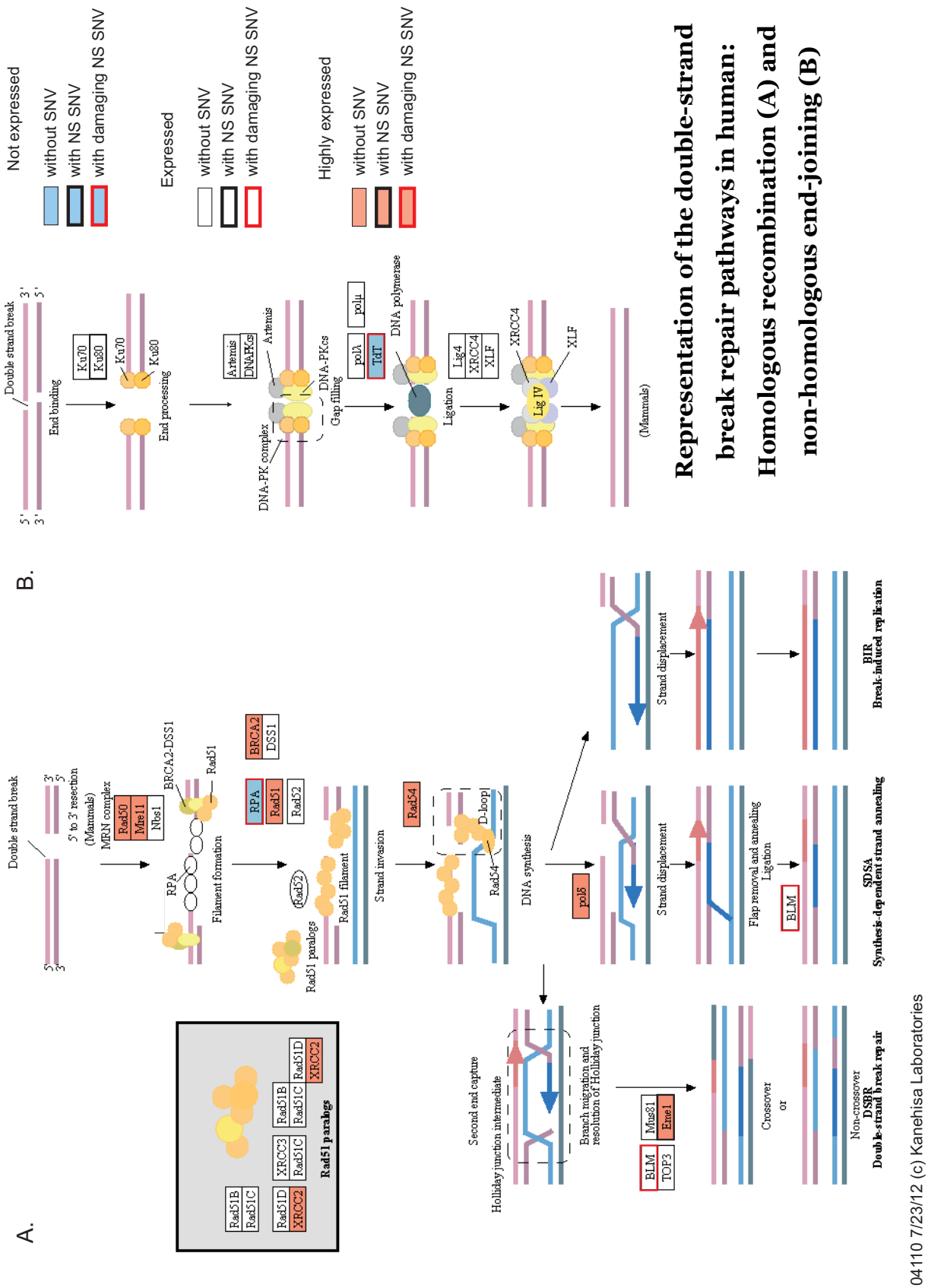
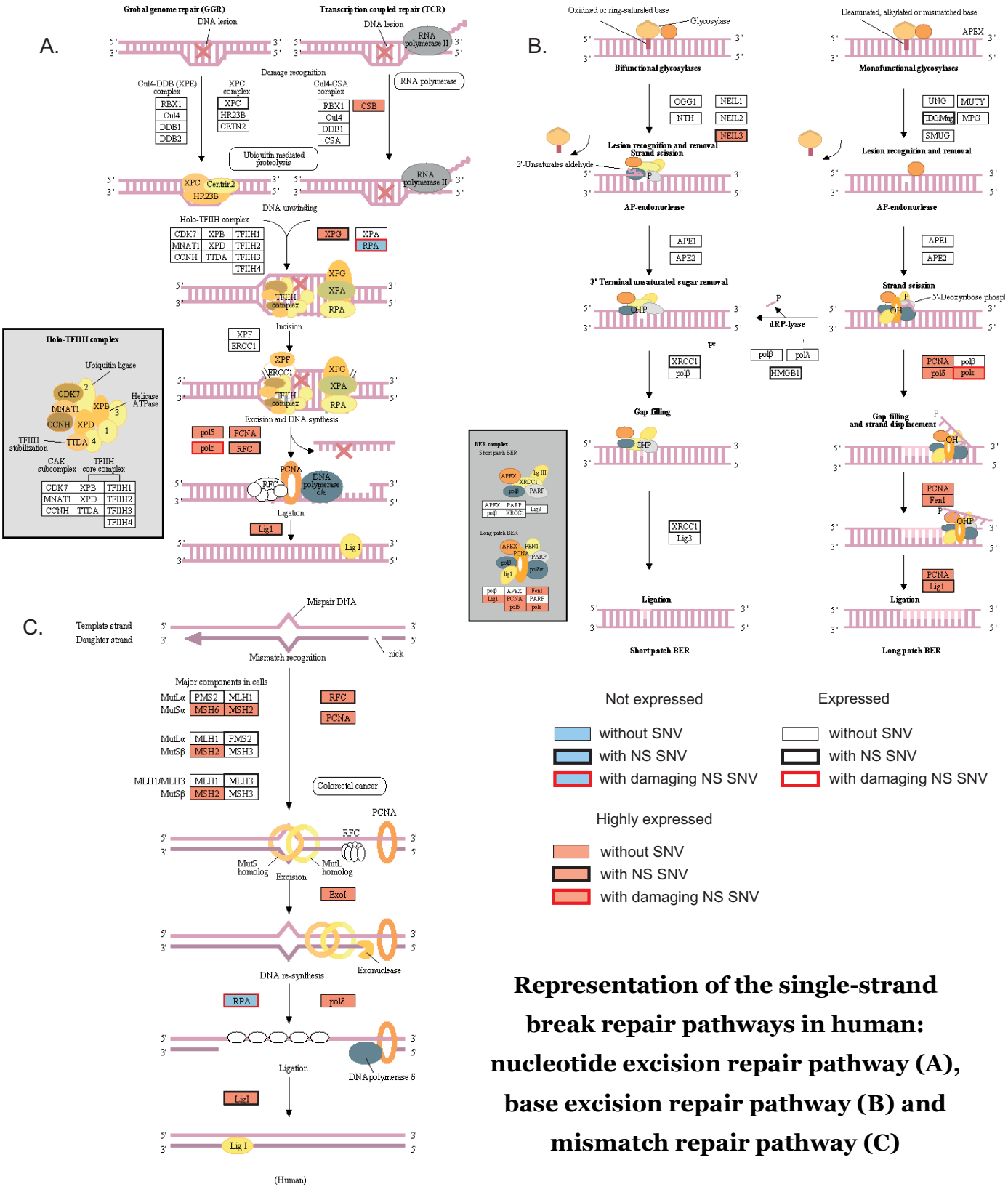


Figure 16: Single strand breaks repair pathways



04110 7/23/12 (c) Kanehisa Laboratories

Representation of the single-strand break repair pathways in human: nucleotide excision repair pathway (A), base excision repair pathway (B) and mismatch repair pathway (C)

4.2.4 Design and interpretation of RNAi reagents

Validating RNAi screen results is one example of the usefulness of the dataset provided in this study. Most RNAi designs, particularly the commercially available ones, are based on the human reference genome [170]. The MitoCheck project developed a high-throughput RNAi screening method [57], which allows the systematic assessment of all genes involved in chromosome segregation and cell division in human (utilizing the same HeLa cell line we sequenced). Out of the 600 genes identified to play a role in mitosis, we took the example of the gene CABP7 (www.mitocheck.org; MCG_0016344), which induced a mitotic defect phenotype that was rescued with mouse transgenes (i.e. a confirmed hit). It was targeted with 4 siRNAs, out of which only 2 induced the defect (MCO_0033274 and MCO_0033284). In the HeLa RNA sequencing data, we found that only MCO_0033274 and MCO_0033284 matched exons from the CABP7 locus, whereas MCO_0056521 and MCO_0056522 (the two non-phenotype-inducing siRNAs) did not match any reads. This is one example of how an absence of phenotype can be explained by non-targeting siRNAs, which was validated by searching for perfect matches between siRNAs and reads. Incorporating HeLa genome sequence information into future design of siRNAs should therefore increase the efficiency of siRNA targeting, yielding more consistent and reliable results.

5 Discussion

By performing deep DNA and RNA sequencing of a HeLa-Kyoto cell line, we provide resources of variation compared to the human genome, including SNVs, SVs, copy number, and additional complex rearrangements such as chromothripsis. We used this information, especially homozygous high-confidence calls, in order to build the first HeLa genome. In addition to the genomic resources, we also profiled the gene expression of this cell line and determined general pathways highly expressed compared to the physiological range of expression across 16 different tissues. In the following sections, these findings and their implications for reevaluate the way HeLa is used to model human biology are discussed. The discussion is focused on the resources provided and their impact, implications of the extensive variations observed in HeLa, and the challenges encountered during this work.

5.1 Resources provided and usability

5.1.1 The HeLa reference genome and accession

In addition to the genome, we also deeply sequenced the transcriptome of the same cell line and give a snapshot of expression profile for this cell line in standard culture conditions. The raw sequencing reads are deposited in the ArrayExpress archive, which is a database of functional genomics experiments[3]. The integration of both types of data enables the possibility of analysing expression level related to genomic sequence variation and especially CN variation. The transcriptome coverage along our HeLa genome is also computed and available[3]. The data can be retrieved and researchers can process them on their own in order to be comparable to their specific expression profiles. By clarifying which genes are expressed in standard conditions, future RNAi screens can be designed taking into account information on which genes are not expressed or not. Indeed such functional classes cannot be well assayed in HeLa.

5.1.2 The HeLa varyome

In parallel, a complete catalog of all HeLa differences compared to the human reference genome including copy number, heterozygous deletions, heterozygous insertions, inversions and interchromosomal translocations are reported and deposited in the EMBL Nucleotide Sequence Database (also known as EMBL-Bank),

which constitutes Europe's primary nucleotide sequence resource under accession numbers available in our web page[3].

5.1.3 Transcriptome

In addition to the genome, we also deeply sequenced the transcriptome of the same cell line and give a snapshot of expression profile for this cell line in standard culture conditions. The raw sequencing reads are deposited in the ArrayExpress archive, which is a database of functional genomics experiments[3]. The integration of both types of data enables the possibility of analyzing expression level related to genomic sequence variation and especially CN variation. The transcriptome coverage along our HeLa genome is also computed and available[3]. The data can be retrieved and researcher can process them in their own in order to be comparable to their specific expression profile. By clarifying which genes are expressed in standard conditions, future RNAi screens can be designed taking into account information on which genes are not expressed in standard conditions. Indeed such functional classes cannot be well assayed in HeLa.

5.2 Variations in HeLa

5.2.1 Extent of variations

To date, these resources represent the most comprehensive genomic characterization of a HeLa cell line. While previous studies have indicated considerable variation between HeLa and the reference genome, this data clearly reveals the extent and nature of this variation. More than 70% of the genome is present in more than 2 copies (Figure 10), and more than 20% present LOH. We identified around 4.5 millions SNVs and around 0.5 million indels in addition to around 3000 SVs including deletions, insertions and interchromosomal translocations (Table 2). These variations affect 1250 genes and the functional impact of these mutations is discussed later.

Over 80% of the identified SNVs and short indels are common variations, which are also present in SNV catalogs such as dbSNP [171] and the 1000 genomes project dataset [113] and which reflect the proportion of the normal short variations exhibited by Henrietta Lacks' genome still present in this cell line. The pool of

variants not present among common variants could be either Henrietta Lacks-specific or HeLa-specific. The Henrietta Lacks' genome would be an advantage to untangle the different groups of variants, but is unlikely to become available to researchers [31].

5.2.2 Why is copy number state 3 favored in HeLa?

Most of the HeLa genome is present in 3 copies. This evolution compared to the normal diploid human state likely reflects the adaptation of HeLa to specific environmental conditions, which could have either taken place during cancer progression or decades of cell culture. Copy number variation has been recently suggested to be an important player in human evolution and adaptation [172]. The authors described how regions exhibiting CN changes are likely to initiate the formation of new genes that could enlarge the pool of genes susceptible to create positive selection in response to environmental changes. An analysis of CN state across multiple tumor types shows a near-triploid CN associated with a high mortality prognosis such as lung, pancreas, gastric and colon cancer [173]. CN 3 seems to favor aggressive tumors and have to some extent a role in fast proliferation and growth for cells grown *in vitro*. This feature could have been a property of the original tumor, described as extremely aggressive, and we can also speculate that it could be one of the reasons why these cells were the first to be successfully cultivated *in vitro*. However the question of how CN 3, and in general aneuploidy, could favor fast proliferation remains unanswered.

5.2.3 Viral insertions

The study presents the first potential set of 10 viral DNA integrations in the HeLa genome. This list compiles, on one hand the initial viral infection (namely the HPV infection), which most likely triggered the development of cervical cancer [174] and on the other hand, probably the result of years of genomic manipulation with integration of diverse viral vectors for experimental purposes such as the Herpes virus for example. The suspected presence of a sequence from the taterapox virus might also be du to genetic manipulation. Indeed this virus is part of the family of the poxviruses virus, which are commonly used in molecular biology to deliver genetic

material in cells (i.e. transduction) and usually contain an origin of replication, a multicloning site, and a selectable marker. A direct consequence of such integration could be the activation of expression of surrounding areas of the genome where the integration sites are localized could potentially impact cellular processes and could have repercussion at a phenotypic level. Instability of such inserts could also allow viral sequence to jump to one site to another and perturb locally the expression pattern of the corresponding genomic area.

5.2.4 M-FISH reveals genomic instability in HeLa

The M-FISH experiment reveals common rearrangements occurring in all cells tested, but it is also important to note that cell-specific events occur to quite a large extent (Figure 12). This observation reflects the great diversity of rearrangements occurring in HeLa and especially single-cell events. We assume that some of these chromosomal aberrations do not cause inviability but this could indicate a constant attempt at DNA shuffling, which could potentially provide a selective advantage for HeLa to respond faster to environmental changes. Although this indicates considerable heterogeneity among the cells sampled for these experiments, we do not observe such heterogeneity in our genomic analysis of CN, for example, indicating that the population of cells sequenced was indeed mostly homogeneous.

5.2.5 HeLa vs. physiological range – lessons learned

Our HeLa transcriptome data shows that only 20% of protein-coding genes are expressed in standard *in vitro* culture conditions. In Human, the Encyclopedia of DNA Elements (ENCODE) project analysed the repertoire of RNAs produced by human cells in order to describe in details functional elements buried in the human genome sequence. Their results [175] describe that the range of expression between the cell type and cell lines (including another HeLa cell line, i.e. HeLa-S3) is quite uniform and that more than 50% of the protein-coding genes are expressed across 14 different human cell lines which is a higher proportion than what we find. The set of expressed genes is likely an essential subset for HeLa to be able to maintain proliferation. In addition, we found that only less than 2000 genes are more highly

expressed than the physiological range of the 16 different tissues from the Illumina BodyMap project (ENA number: ERP000546). The GO term enriched in this subset of genes are related to proliferation, transcription and DNA repair. The last term might indicate that even though HeLa displays a high level of chromosomal instability, specific DNA repair mechanisms are trying to compensate for the perturbations and minimize the damage. A closer examination of these pathways by integrating transcriptomic data and sequence variation data could highlight players involved and give rise to new insights of DNA repair mechanism. This aspect will be discussed in more detail in the next section.

RNA sequencing protocol and analysis is still challenging. The library preparation contains steps, which can affect the analysis of those data. Regarding gene expression assessment, in theory, using a gene model and a set of mapped reads, the read counts can be summed up for each gene and the expression level can be deducted for every genes at this sequencing depth. The number of reads falling into gene exons is a function of the length of the mRNA as well as the quantity of mRNA molecules present in the sample [176]. The reliability of the quantification of gene expression is correlated to the read mapping performance. The human genome contains repetitive and homologous regions, which cause mapping ambiguities and also add an additional complexity.

5.2.6 Predicted non-functional proteins and impact on HeLa biology

What are the functional implications of the extensive sequence variation we observe in the HeLa genome? Normal cells maintain and tightly control the availability of growth-inducing molecules, which drive the entry into and progression through the cell cycle. The deregulation of these controls is common in tumorigenesis [140]. Globally, the players involved in the cell cycle network are mostly expressed at higher level than the physiological range of expression (when comparing to the BodyMap data). In addition, more than a dozen genes involved in the cell cycle contain damaging NS mutations and are not expressed in HeLa, indicating inactivation of these players. Most of these genes are negative regulators for entry into different cell cycle phases, such as mitosis (WEE [162]), S phase (E2F ORC, MCM and CDC7 [163]), G1 phase (P19 [164]). More global regulators are also part of this group,

including P53, P300, GADD45, TGFβ [165] and 14.3.3 protein complexes [166]). Altogether, these data suggest a potentially uncontrolled activation of the cell cycle leading to constant proliferation.

Normally P53 integrates inputs that evaluate genomic damage and stress conditions and subsequently trigger apoptosis if the conditions are not optimal. P53 mutations are reported to happen in almost every type of cancer at varying frequencies (from 6% in cervix to 42% for specific ovarian cancer - IARC P53; release R15; [167]) and it is likely that these mutations and perhaps others in this pathway can be attributed to the tumor development *in vivo*. As observed in HeLa, a weaker P53 activity, due to an NS mutation, leads to inappropriate replication of cells, consequently allowing cells with genomic damages or inappropriate DNA content to replicate. In addition to this P53 alteration, the genes involved in DNA repair pathways are highly expressed compared to the physiological range of expression (Table 3), however each of these pathways contain at least one predicted non-functional gene (Figure 15 and Figure 16). These observations suggest that HeLa is still able to sense the damage occurring in its DNA and tries to fix it by activating all dedicated pathways but they remain inefficient due to damaged components. Another possible explanation would be that the pathways are weakly preventing even worse damage from occurring.

Over 1000 non-synonymous SNVs are predicted by SIFT to affect protein functionality and an enrichment of those SNVs in genes categorized in “sensory perception of chemical stimulus” probably reflects the deregulation of specific receptors and the constitutive activation of signaling pathway players operating downstream of these receptors. Mutations in negative-feedback mechanisms, which attenuate growth signaling, are therefore capable of enhancing proliferation. We observe in HeLa that an important integrator of such signals carries NS mutations in its sequence: the Ras oncoprotein. Mutations can affect its GTPase activity and then disrupt this negative-feedback mechanism resulting in constant activation of the signal transmission.

Aside from the cell cycle and by integrating homozygous small indels and large deletions and insertions into the human genome, exons of 14 genes were disrupted. The gene group is formed of 5 pseudogenes, 3 RNA genes and 6 protein-coding genes. One of the genes affected is the alpha-tubulin (TUBA4B) protein. It is potentially a component of microtubules, which forms the spindle fibers for

separating chromosomes during mitosis. The disruption of such genes could, to some degree, affect missegregation of DNA material during this process.

5.2.7 Chromothripsis and development of cancer

Complex chromosomal rearrangements, referred to as chromothripsis [85], have recently been associated with 2–3% of all cancers, and specific examples have been described in many different cancer types [136,177-180]. This rather common phenomenon in cancer is also observed in HeLa. Two different interpretations of this observation could be made.

The first view would consider chromothripsis as a cancer catalyst [137]. The impact of such a catastrophic single event is potentially far higher for the cell than the common progressive model of tumor development, which involves sequential mutational events randomly throughout the genome over time. The result of this large pattern of rearrangements or mutations might have triggered carcinogenesis pathways and resulted in Henrietta Lacks' cancer. Another hypothesis would suggest that chromothripsis might have occurred over the years in cell culture and could potentially create new rearrangements *in vitro*.

A mechanism by which chromothripsis arises [105] involving mitotic chromosome missegregation has been proposed. The Smc 1, Smc3, Separin, Securin, Polo protein family, Aurora B proteins, BubR1, Cdc14 and the anaphase-promoting complex/cyclosome (APC/C) protein complex are involved in cell spindle formation, the regulation of chromosome binding to spindle microtubules, guarantee spindle checkpoints, and enable daughter cells to achieve cytokinesis [181-184]. In HeLa, predicted NS SNVs and damaging NS SNVs have been localized in all the genes encoding the previously mentioned proteins, which could create anomalies during mitosis and a favorable environment for chromothripsis. As *in vivo* and in this context, the occurrence of chromothripsis could be seen as an evolutionary benefit for tumor-derived cell lines. It is hard to distinguish between the different hypotheses since we do not have access to Henrietta Lack's genome sequence.

Independently of the origin of the phenomenon, there are clear indications of chromothripsis occurring on chromosome 11 in HeLa. This genomic interval also overlaps with regions presenting LOH and where tumor suppressor gene(s) have been predicted (11q13). These observations support a relationship between complex

chromosomal rearrangements and factors favoring cell development toward cancer phenotype and rapid proliferation.

5.2.8 Loss of heterozygosity regions and cancer

From our study, genomic instability is characteristic of the HeLa cell line. Aneuploidy is often caused by errors of chromosome segregation during mitosis. Some tumor cells are stably aneuploid and proliferate normally [185], however aneuploidy often results from chromosomal instability characterized by an increase in the rate of gain or loss of DNA material during cell division [186], using uncharacterized mechanisms.

Some potential indications have been developed on how aneuploidy could have a role on genomic instability and mentioned three different directions [187]: aneuploidy could create protein disproportions and therefore stimulating genomic instability. A second hypothesis is that aneuploidy could promote loss of heterozygosity by duplicating DNA segments containing oncogenic alleles or allowing for the loss of the remaining wild-type copy of a tumor suppressor gene. An alternative explanation favors aneuploid cells to protect their DNA from deleterious mutations in essential and haplosufficient genes by allowing the cell to survive longer in order to accumulate crucial growth- promoting and transforming mutations. Overall, in HeLa, 23% of the genome exhibits LOH and therefore ¼ of the genome carries only one gene version, which can thus be easily impacted by disrupting events.

5.2.9 Dosage compensation

CN is expected to impact gene expression levels in a proportional manner unless dosage compensation occurs [188,189]. This mechanism plays an important role in *Drosophila melanogaster*. In this model and for autosomal genes, the expression of heterozygous loci is twofold less than the homozygous reference loci expression [190]. Our result shows, that in the genomic context of CN 3, the allelic ratio 2:1 is conserved and therefore either there is no dosage compensation or the compensation impacts all alleles equally and not very strongly. This finding also corroborates observations by Schlattl et al. [191], who evaluated the impact of CNVs on transcript

level across 129 individuals and found only 3 genes out of the 12 tested, possibly affected by dosage compensation.
Potential lack of dosage compensation could contribute to protein imbalance observed in other aneuploidy cells in human [192]. It is proposed that cells respond to this protein overload by activation of autophagy to correct protein levels. The observation that genes exhibiting higher level of expression in HeLa are enriched in the GO term proteolysis could corroborate this hypothesis.

5.3 What are the data useful for

5.3.1 Use of the HeLa genome for molecular studies instead of the human reference

When working with HeLa cells, scientists are currently limited to the use of the human reference genome since they lack more relevant sequence information about their cell line. This set of resources bridges the gap and provides for the first time a HeLa genome sequence available to anyone who requires specific sequence information in order to get insights about specific genes or the design of any experiment. The integration of high confidence sequence variation calls into the human reference genome constitutes a better representation of the genome of a HeLa cell.

5.3.2 RNAi design and interpretation

RNAi-based technology have been largely used with mammalian cells for studying gene function [193] and HeLa cells is frequently applied with these methods [54,57,58,194-202].
Despite the publication of RNAi screen design guidance rules [203,204], large-scale RNAi screens tend to exhibit relatively high rates of false positives due to off-target effects and the possible reasons are still debated in the literature [205,206]. Commercial genome-wide RNAi libraries are designed based on the human reference genome and therefore the probes can overlap genomic variations between the two genomes. An example of such off-target effects has been discussed for the CABP7 gene in our result section. The use of the HeLa genome sequence will enable better probe design and potentially reduce off-target effect.

The availability of genomic and transcriptomic data for HeLa might be useful to reinterpret data already produced using the cell line and could help anyone for the RNAi design for further experiments. This could apply to a larger extent to any model used for cell-based RNAi screens and filter any sequence variation, which could represent a potential off-target source.

In HeLa, another layer of complexity for analyzing the data produced by RNAi screens is the CN variation along its genome. Despite the fact that we did not observe a general trend for dosage compensation in the expression levels of CN 3, 4 and 5 regions, expression of certain genes may not be affected by CN variation. In the context of expression knockdown mediated by RNAi, the interpretation of expression value for a gene present more than twice in the genome has to be made carefully depending, for example, on the efficiency of the silencing for each copy by the RNAi probe.

5.3.3 Recommendations for future studies involving HeLa

Based on our observations in terms of genomic complexity and the narrow subset of genes still operating in HeLa, future molecular studies using this cell line, independently of their scale, should be carefully designed regarding the biological question asked, the experimental design and the interpretation of the results. Any experiment involving gene expression quantification should take into account CN for each specific gene. Genetic manipulation (gene replacement, integration) should be carefully validated to avoid incomplete penetrance of the genomic sequence, for example a case where a gene would have been replaced or deleted in 2 out of the 3 copies of the original sequence.

5.3.4 General comments on studies using cell lines

Since 1952 and the establishment of the HeLa cell line, working with cell lines in research has offered a unique advantage in providing an infinite stock of cells, which continuously proliferate under standard *in vitro* conditions.

Despite the extensive use of this particular cell line over the years, phenotypic and genotypic differences arise between clonal populations originating from the same cell

line. Different explanations for this phenomenon can be found. Firstly, cell lines naturally generate variants exhibiting phenotypic and genotypic differences from the original clone. If one of these subclones is a fast-grower, it will overtake the cell culture and will represent the largest proportion of the cell culture. Secondly, cells are subjected to different environmental conditions, which exert selection pressure on the variants and favor a subpopulation of variants better adapted to the environment. This mechanism has probably occurred for several decades and given rise to very different HeLa cell sublines. This is potentially also true for other cell lines grown indeterminately and passed between laboratories, which lost phenotypic and genomic traits from the original tumor. Those divergences can be exacerbated by genomic instability, as observed in HeLa and could likely affects other cancer cell lines due to the genomic features that they need to acquire in order to become tumoral.

Cell lines can nonetheless remain very useful in the future given specific cell culture rules (for example limited number of passages, frozen cell stocks [207]). Based on our characterization of the HeLa genome, it seems that extra-precautions should be made when using any cell line and a full characterization at regular intervals would be recommended to ensure robust data and reliable results.

Cell lines played an indispensible role in our understanding of biology and are still used by researchers to model human cells or human disease. They will become even more necessary in current efforts to bridge the gap between research and individualized therapies [208]. The production and the maintenance of those cell pools from individuals must be very standardized and controlled to avoid any phenotypic and/or genotypic changes from the native source. The recent development of induced pluripotent stem (iPS) cells could give a new impetus cell lines. iPS are the result of somatic cell reprogramming to an embryonic-like state with the potentiality to give rise to all differentiated cell lineages. This property seems to be associated with the epigenetic state of the cell comparable to embryonic stem cells [209]. Methods to achieve the transformation use transcription factors cocktails [209-213]. Among the possible applications of iPS [214], these cells can provide models for normal development , for cancer research and for human disease,. iPS-cell technology opens the production of cell lines from patients with various diseases [215]. Another application of iPS cells could be to model normal development in humans, which could help to understand disease mechanism and to

initiate and develop therapies in some cases [216]. The area of cancer research could also be impacted by the emergence of iPS cells. They can be derived from tumor cells in order to assess the importance of the genetic background of an individual developing the disease. iPS, therefore could be seen as the future of cell lines and could also benefit from the knowledge, established with “old” cell lines like HeLa, on the good practice of cell culture.

5.4 Assembling an aberrant genome

In order to characterize a genome, originated from human tumor cells, we performed deep sequencing and used the reads to establish a list of variants compare to the human reference genome. The standard approach that we used is to map the reads against the reference and extract positions where HeLa contains simple variant sequences. The fact that this HeLa genome contains a different copy number for most of its DNA segments complicate the analysis by adding another layer of complexity in order to reconstruct all haplotypes without any DNA information from Henrietta Lacks.

Assembly could be potentially applied to short read sequences to align and overlap the DNA fragments of longer DNA distance to reconstruct the original sequence. The major problem for applying this approach is that DNA sequence reads may fit together in more than one way because of repetitive sequences within the genome. As the fraction of unique sequences increases with the read length, assembly would be a method of choice when eventually every sequence in the genome is unique.

However, although there are established algorithms for de novo assembly from HTS shotgun data, which are based on overlap [217] or de Bruijn graphs [218-220], current approaches have limitations. Current assembly methods also typically ignore pre-existing information, such as a reference sequence or known variants. New development in *de novo* assembly algorithms, and especially an approach called Cortex [221], focused on detecting and characterizing genetic variation in one or more samples. This method, which extends classical de Bruijn graphs, could accommodate information from multiple samples, including one or more reference sequences and known variants and was already employed by the 1000 Genomes Project [113]. This new alternative approach might be tested on our dataset to try to reconstruct in more detail the genome of this cell line.

6 Conclusion

While HeLa is one of the most widely used model cell line for studying human cellular and molecular biology, a comprehensive characterization of its genome has not been described so far. Low-resolution cytogenetic approaches have highlighted the importance of the genetic rearrangements occurring in those cells [75]. We performed deep DNA and RNA sequencing in order to assess the mutational portfolio and the gene expression profile of a HeLa Kyoto cell line. We built the first HeLa genome and provide a list of resources directly usable for the community. In addition, the lists of SNVs, deletions, inversions and tandem duplications are made available for the community as well as the copy number along the HeLa genome. This set of information could be downloaded from public database repository for which all accession numbers can be found at this link[3].

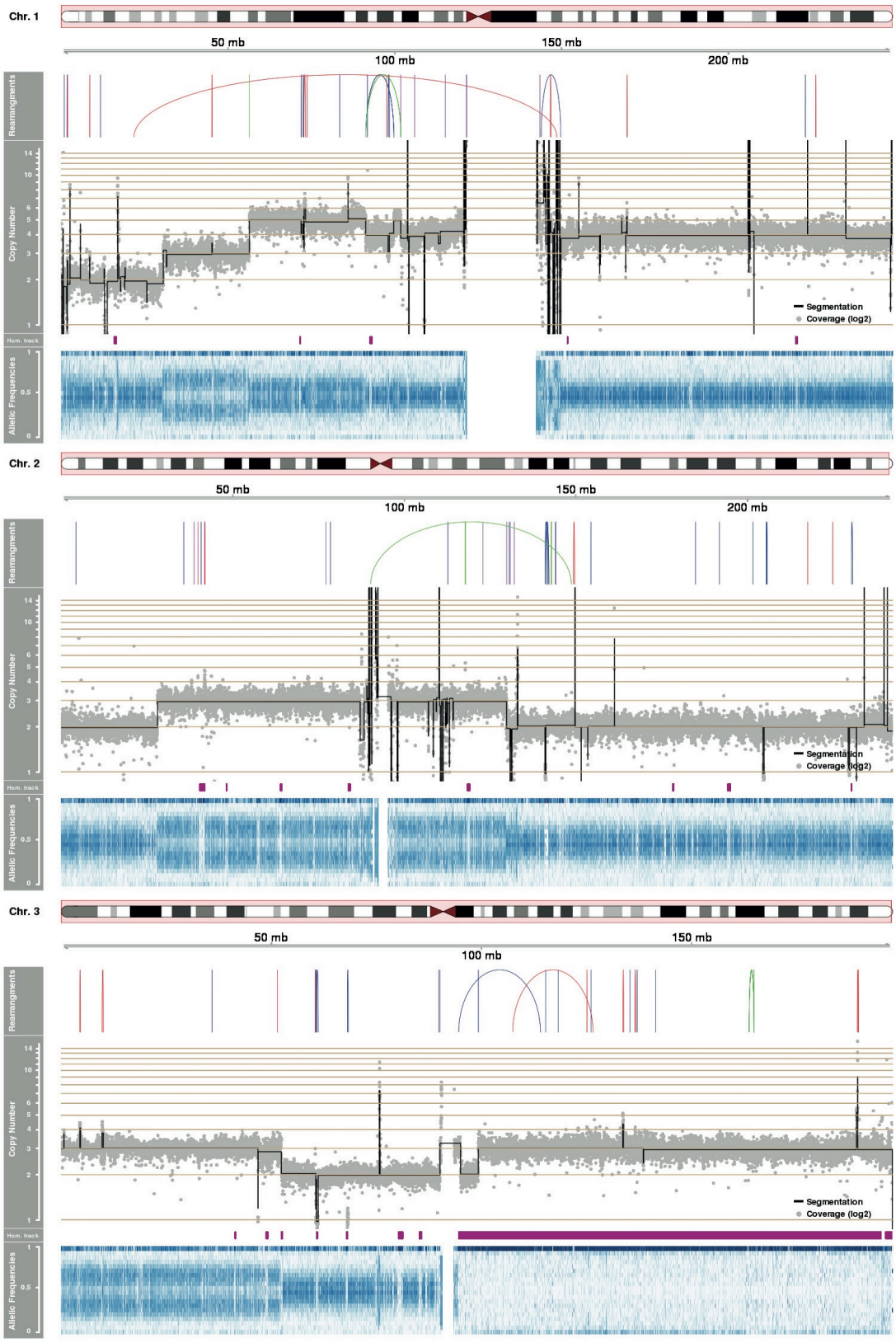
In summary, more than 40 % of the genome is in 3 copies (Figure 9) and it contains a set of complex structural variants (Figure 10) including deletions, tandem duplications, inversions and translocations. We also emphasize the strong indications for the occurrence of chromothripsis (Figure 11) in the cell line. The multiple rearrangements observed during this analysis corroborate an extreme genomic instability, which is one of the main characteristics of cancer cells. We also present an integrative view combining SNVs predicted to impact directly the protein function, overlaid by gene expression levels, which give rise to numerous hypotheses on Hela cellular behavior.

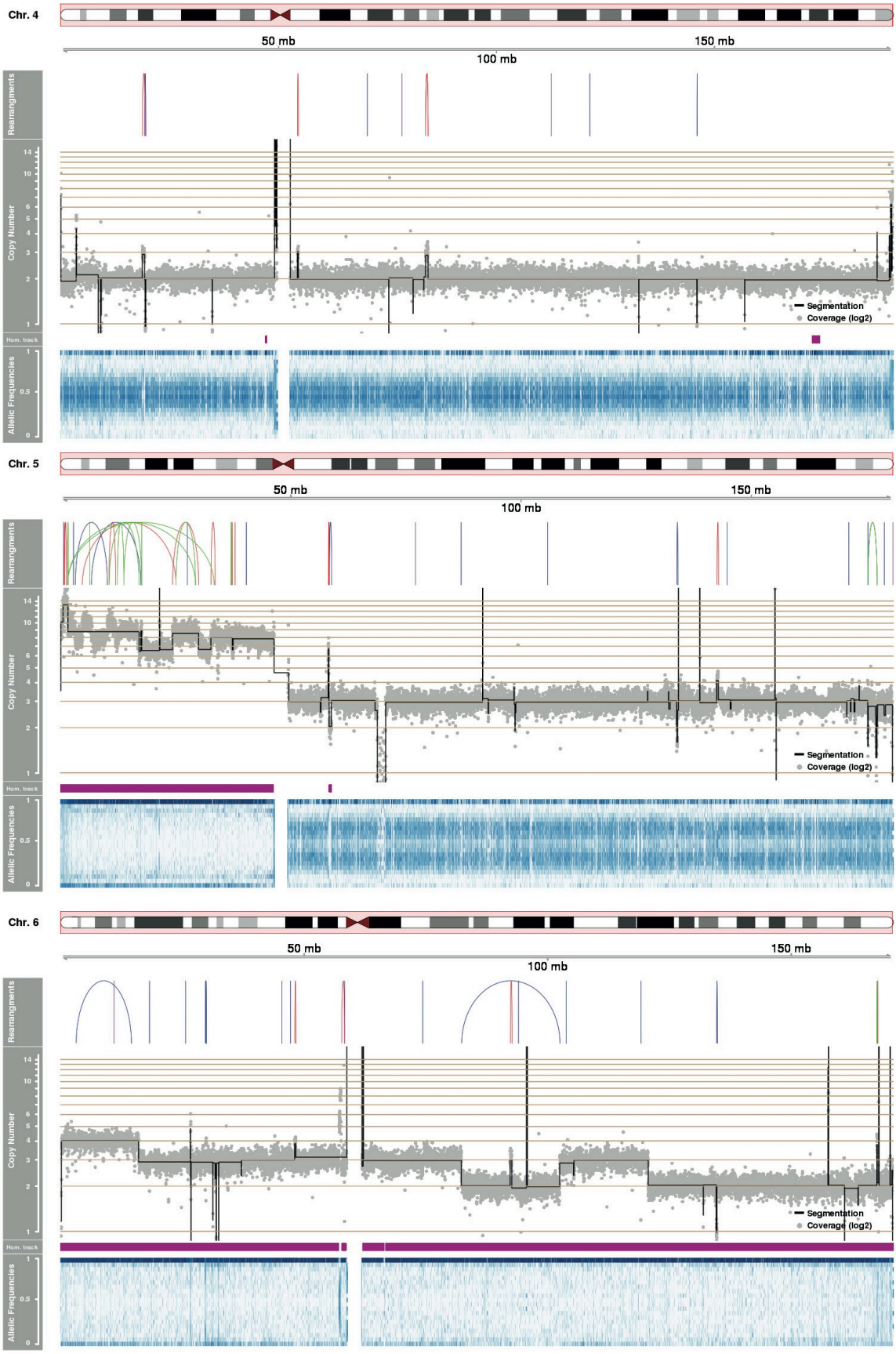
We investigate, in particular, the different DNA repair pathways and propose that despite an important genomic instability, which often leads to the occurrence of structural variants, HeLa is still able to sense the DNA damage and to activate gene expression for the players involved in DNA repair pathways (Figure 14, Figure 15 and Figure 16). Inside those pathways, we notice at least one player predicted to be non-functional and therefore postulate that despite the activation of such pathways, they remain inefficient due to defective components. As a consequence of this inability to fix DNA abnormalities, severe sequence rearrangements could occur at a single cell level, as detected with M-FISH experiments. Nevertheless those events would confer specific growth advantage and therefore the appearance of subclonal populations, which is in opposition to what is perceived in the DNA segmentation analysis. Indeed while attributing a copy number state for each fragment of the genome, we could not isolate intermediate copy number states, confirming the homogeneity of the cell population.

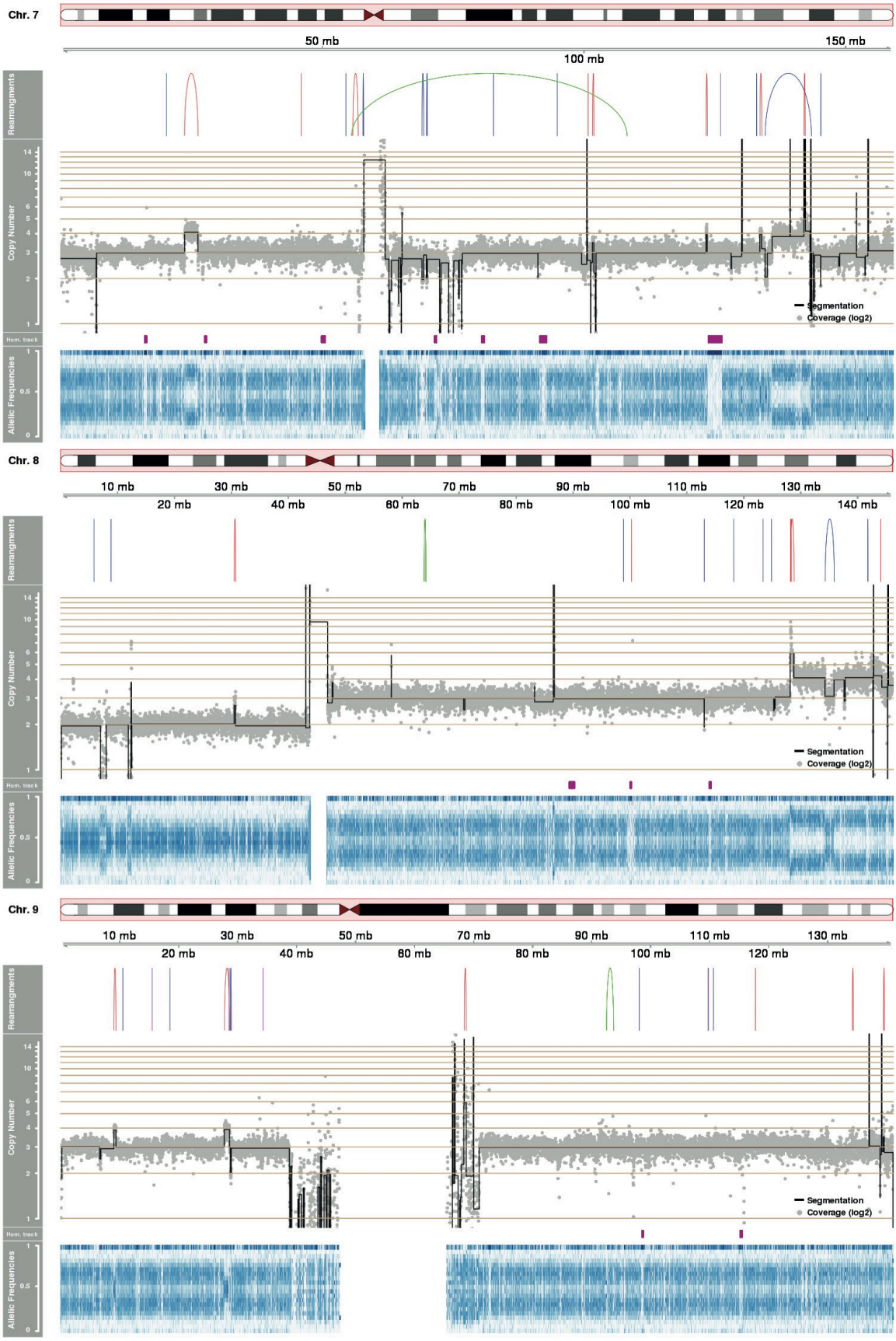
The comprehensive view of this genome combined with transcription data (in standard conditions) enables to understand the biology underlying its phenotypic characteristics and the resource provided by this work could be a valuable resource to assess the genomic content or the standard expression profile of this cell line. The domains of application could be very wide, starting from low-throughput experiments of genes of interest to large-scale approaches such as genome-wide RNAi or proteomic screens, through analysis of specific pathways using HeLa cells.
This is probably one of the first steps to catalog the differences between the human reference genome and the HeLa genome. In the future, the development of short-read assembly technology might provide an alternative option to resolve in an even higher resolution to approach better the complexity of this genome.

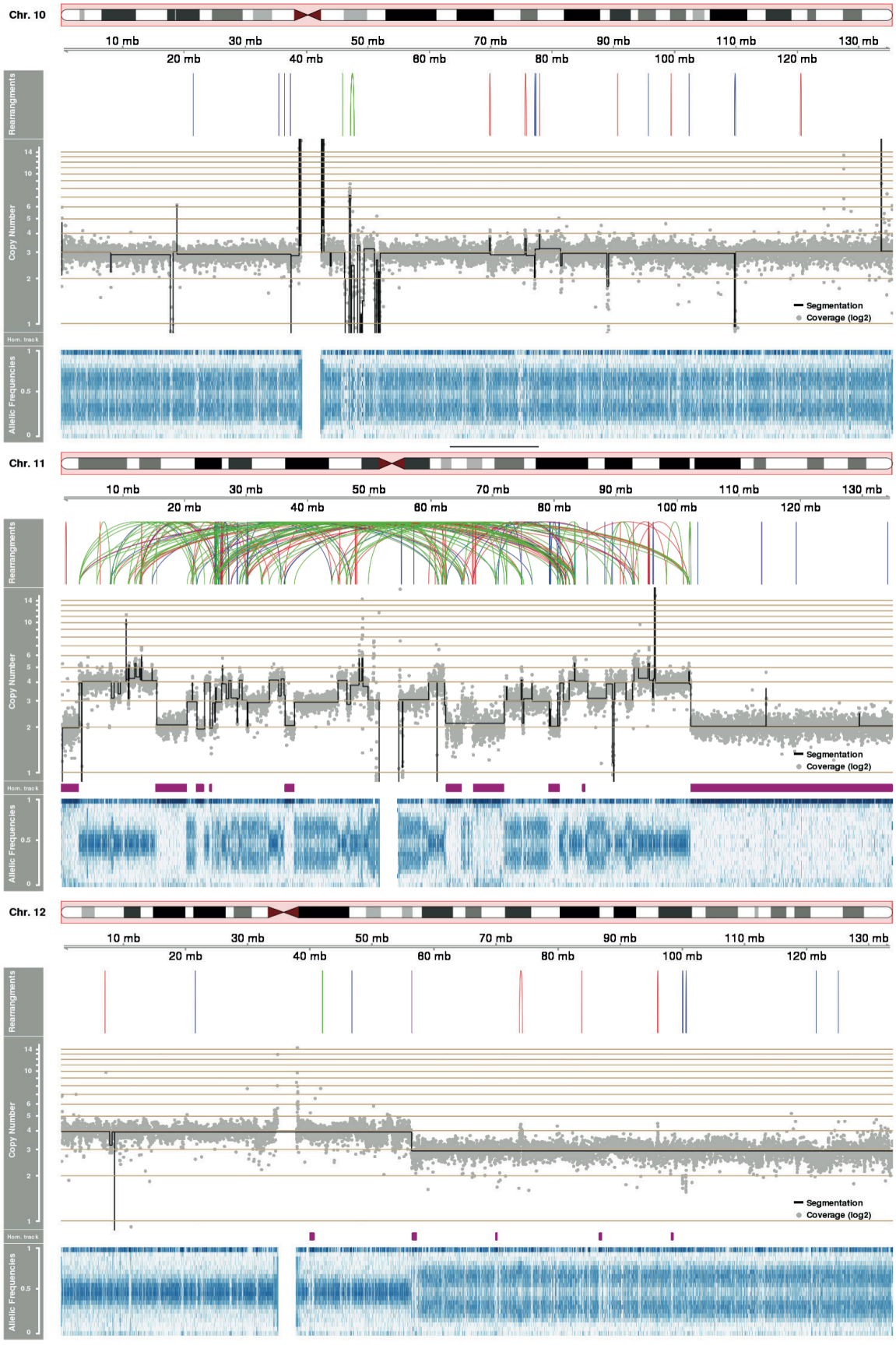
7 Appendix

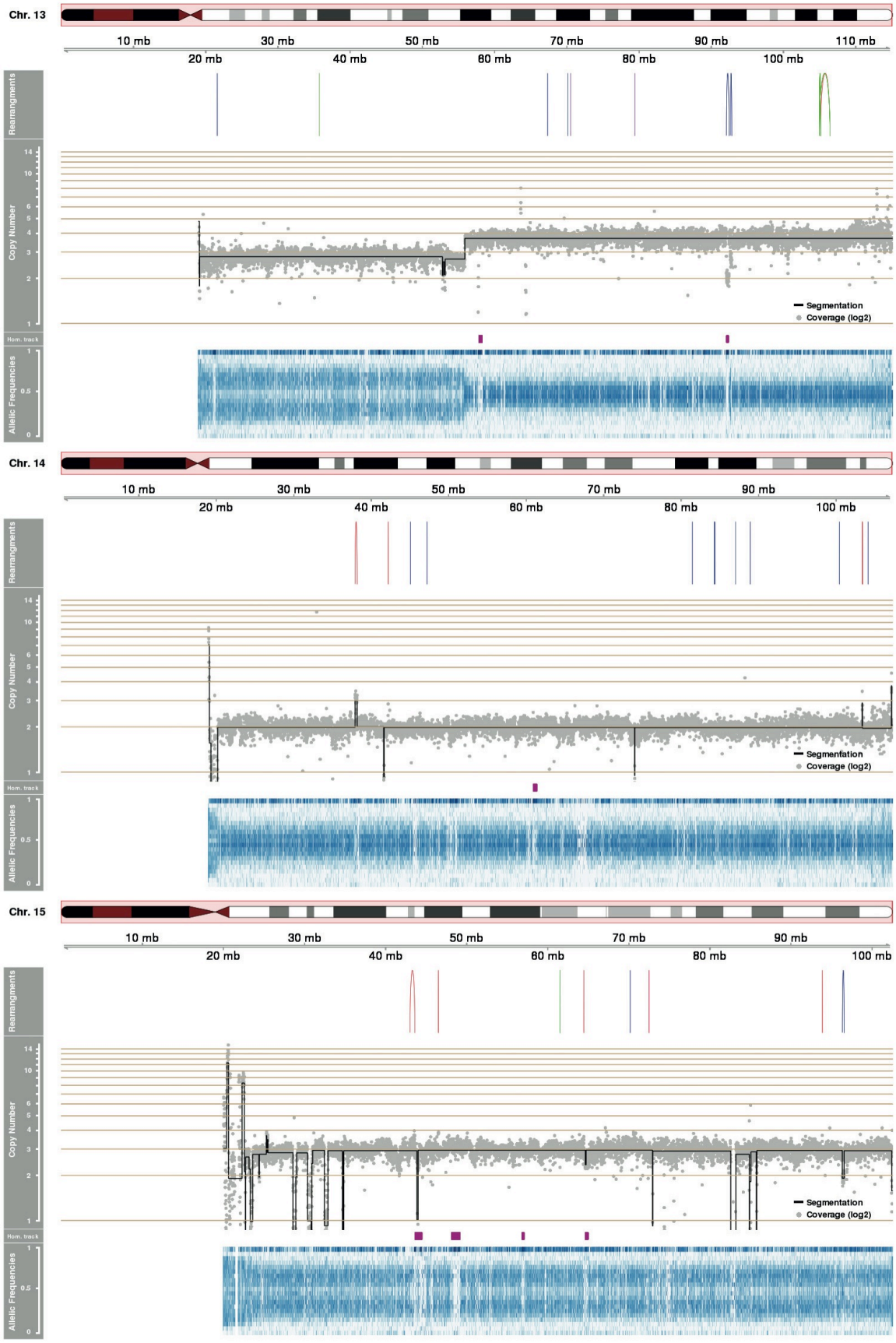
Appendix A: SVs, CN, LOH along chromosomes

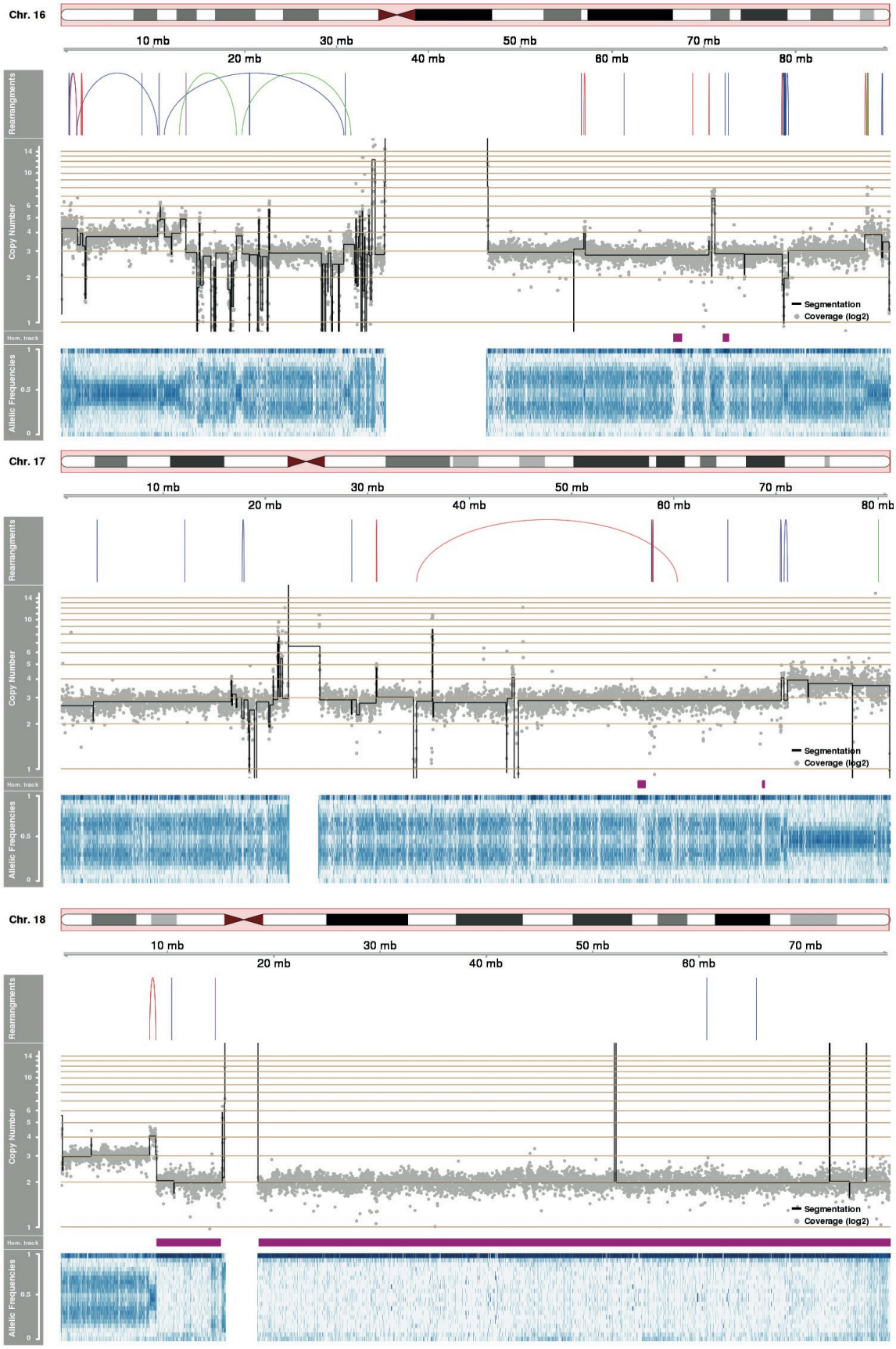


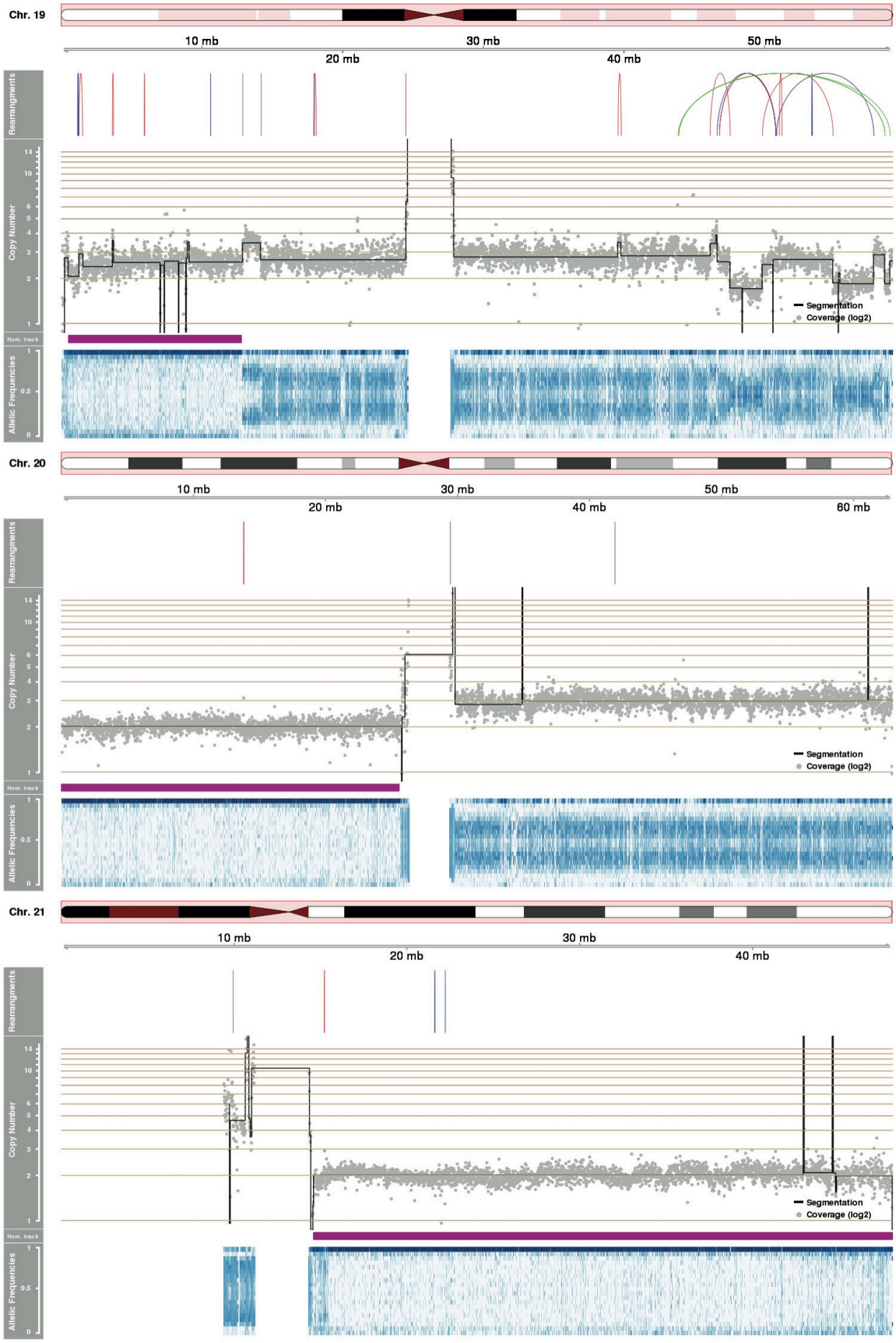

Chr. 4
50 mb
100 mb
150 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies
Chr. 5
50 mb
100 mb
150 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies
Chr. 6
50 mb
100 mb
150 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies

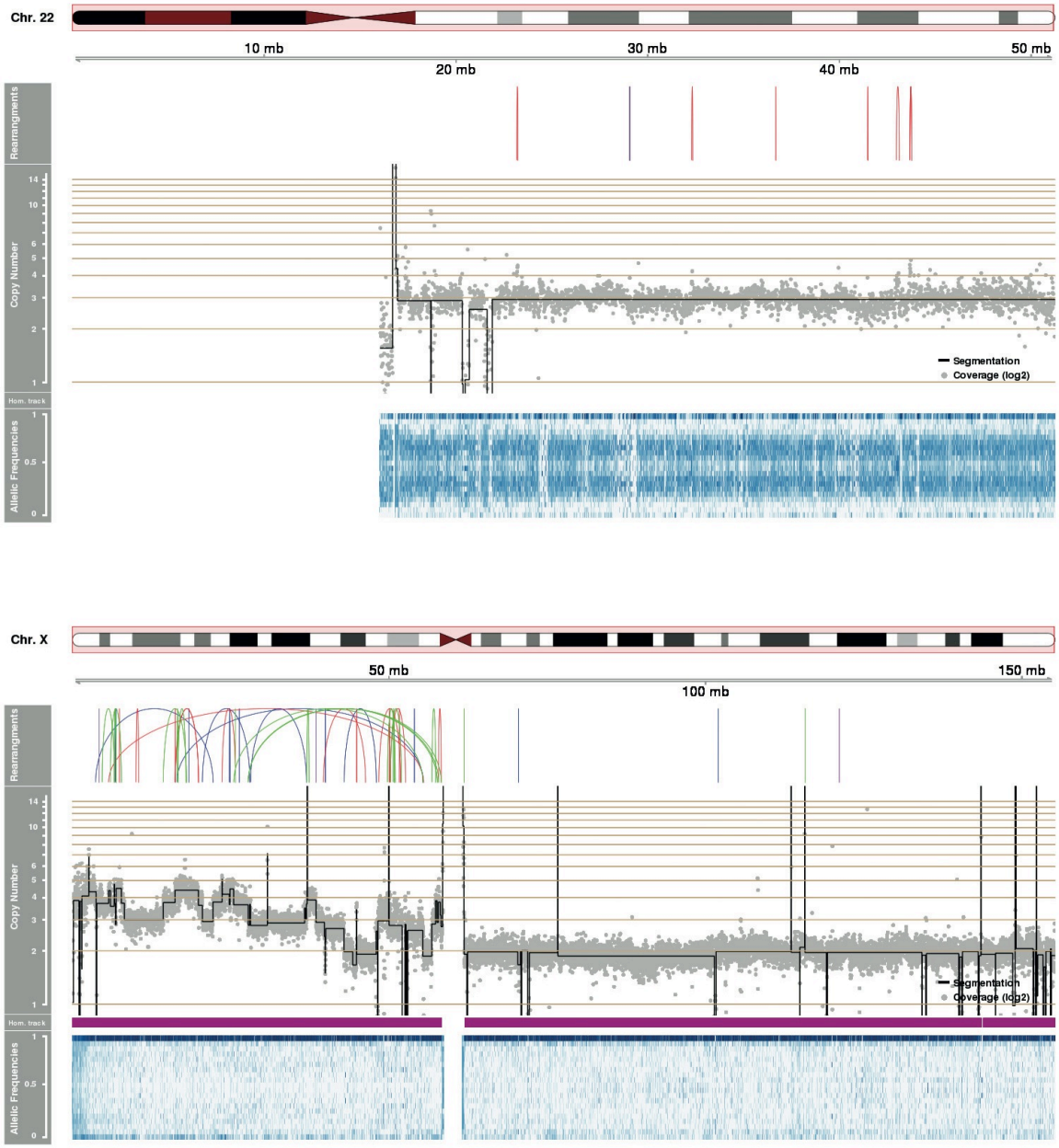

Chr. 7
50 mb
100 mb
150 mb
Rearrangments
Copy Number
Hom. track
Allelic Frequencies
Segmentation
Coverage (log2)
Chr. 8
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
90 mb
100 mb
110 mb
120 mb
130 mb
140 mb
Rearrangments
Copy Number
Hom. track
Allelic Frequencies
Segmentation
Coverage (log2)
Chr. 9
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
90 mb
100 mb
110 mb
120 mb
130 mb
Rearrangments
Copy Number
Hom. track
Allelic Frequencies
Segmentation
Coverage (log2)


Chr. 10
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
90 mb
100 mb
110 mb
120 mb
130 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies
Chr. 11
Chr. 12


Chr. 13
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
90 mb
100 mb
110 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies
Chr. 14
Segmentation
Coverage (log2)
Chr. 15
Segmentation
Coverage (log2)


Chr. 16
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies
Chr. 17
Chr. 18


Chr. 19
10 mb
20 mb
30 mb
40 mb
50 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies
Chr. 20
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies
Chr. 21
10 mb
20 mb
30 mb
40 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies


Chr. 22
10 mb
20 mb
30 mb
40 mb
50 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies
Chr. X
50 mb
100 mb
150 mb
Rearrangments
Copy Number
Segmentation
Coverage (log2)
Hom. track
Allelic Frequencies

Appendix B: Viral insertions

Chr	Start	End	Strand	Read support	VirusType	Virus mapped
1	10002	10118	+	12	herpesvirus	Human herpesvirus 6B, complete genome
1	10002	10118	+	12	herpesvirus	Equid herpesvirus 2, complete genome
1	10002	10118	+	12	herpesvirus	Human herpesvirus 6A, complete genome
1	10002	10118	+	12	herpesvirus	Human herpesvirus 7, complete genome
1	10002	10118	+	12	herpesvirus	Gallid herpesvirus 2, complete genome
1	10002	10118	+	12	herpesvirus	Gallid herpesvirus 3, complete genome
1	10002	10118	+	12	herpesvirus	Meleagrid herpesvirus 1, complete genome
1	10002	10118	+	12	herpesvirus	Ovine herpesvirus 2, complete genome
1	10002	10118	+	12	herpesvirus	Cyprinid herpesvirus 3, complete genome
1	10002	10118	+	12	herpesvirus	Saimiriine herpesvirus 1, complete genome
8	128189764	128190032	-	552	papillomavirus	Human papillomavirus - 18, complete genome
8	128189764	128190032	-	552	papillomavirus	Human papillomavirus type 32, complete genome
8	128192272	128192499	+	23	papillomavirus	Human papillomavirus - 18, complete genome
8	128193167	128193435	+	174	papillomavirus	Human papillomavirus - 18, complete genome
8	128200419	128200690	+	163	papillomavirus	Human papillomavirus - 18, complete genome
12	49659073	49659136	-	6	herpesvirus	Human herpesvirus 5, complete genome

Chr	Start	End	Strand	Read support	VirusType	Virus mapped
12	95467	95567	+	9	herpesvirus	Human herpesvirus 6B, complete genome
12	95467	95567	+	9	herpesvirus	Human herpesvirus 6A, complete genome
12	95467	95567	+	9	herpesvirus	Human herpesvirus 7, complete genome
12	95467	95567	+	9	herpesvirus	Gallid herpesvirus 2, complete genome
12	95467	95567	+	9	herpesvirus	Gallid herpesvirus 3, complete genome
12	95467	95567	+	9	herpesvirus	Meleagrid herpesvirus 1, complete genome
12	95467	95567	+	9	herpesvirus	Cyprinid herpesvirus 3, complete genome
13	19648667	19648780	+	7	poxvirus	Taterapox virus, complete genome
13	19649043	19649209	-	8	poxvirus	Taterapox virus, complete genome
X	155185203	155185362	-	8	herpesvirus	Human herpesvirus 6B, complete genome
X	155185203	155185362	-	8	herpesvirus	Equid herpesvirus 2, complete genome
X	155185203	155185362	-	8	herpesvirus	Human herpesvirus 6A, complete genome
X	155185203	155185362	-	8	herpesvirus	Human herpesvirus 7, complete genome
X	155185203	155185362	-	8	herpesvirus	Gallid herpesvirus 2, complete genome
X	155185203	155185362	-	8	herpesvirus	Gallid herpesvirus 3, complete genome
X	155185203	155185362	-	8	herpesvirus	Meleagrid herpesvirus 1, complete genome
X	155185203	155185362	-	8	herpesvirus	Ovine herpesvirus 2, complete genome
X	155185203	155185362	-	8	herpesvirus	Cyprinid herpesvirus 3, complete genome
X	155185203	155185362	-	8	herpesvirus	Saimiriine herpesvirus 1, complete genome

8 References

Accession numbers are available on our website:
http://steinmetzlab.embl.de/helaseq/
Username: helaseq
Password: helaseq

1. Jones HW (1997) Record of the first physician to see Henrietta Lacks at the Johns Hopkins Hospital: history of the beginning of the HeLa cell line. Am J Obstet Gynecol 176: S227–S228.

2. Lucey BP, Nelson-Rees WA, Hutchins GM (2009) Henrietta Lacks, HeLa cells, and cell culture contamination. doi:10.1043/1543-2165-133.9.1463.

3. Meisels A, Fortin R (1976) Condylomatous lesions of the cervix and vagina. I. Cytologic patterns. Acta Cytol 20: 505–509.

4. Boshart M, Gissmann L, Ikenberg H, Kleinheinz A, Scheurlen W, et al. (1984) A new type of papillomavirus DNA, its presence in genital cancer biopsies and in cell lines derived from cervical cancer. EMBO J 3: 1151–1157.

5. Dürst M, Gissmann L, Ikenberg H, Hausen zur H (1983) A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. Proc Natl Acad Sci USA 80: 3812–3815.

6. Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, et al. (1999) Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol 189: 12–19. doi:10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F.

7. Moody CA, Laimins LA (2010) Human papillomavirus oncoproteins: pathways to transformation. Nat Rev Cancer 10: 550–560. doi:10.1038/nrc2886.

8. Torsvik A, Røsland GV, Bjerkvig R (2012) Spontaneous transformation of stem cells in vitro and the issue of cross-contamination. Int J Biol Sci 8: 1051–1052. doi:10.7150/ijbs.3665.

9. Duensing S, Münger K (2004) Mechanisms of genomic instability in human cancer: insights from studies with human papillomavirus oncoproteins. Int J Cancer 109: 157–162. doi:10.1002/ijc.11691.

10. Shen L (2012) In reply to torsvik et Al. "Spontaneous transformation of stem cells in vitro and the issue of cross-contamination". Int J Biol Sci 8: 1053–1054. doi:10.7150/ijbs.8.1053.

11. Hausen zur H (2002) Papillomaviruses and cancer: from basic studies to clinical application. Nat Rev Cancer 2: 342–350. doi:10.1038/nrc798.

12. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. Cell 100: 57–70.

13. Arruda S, Bomfim G, Knights R, Huima-Byron T, Riley LW (1993) Cloning of an M. tuberculosis DNA fragment associated with entry and survival inside cells. Science 261: 1454–1457.

14. Artandi SE, DePinho RA (2010) Telomeres and telomerase in cancer. Carcinogenesis 31: 9–18. doi:10.1093/carcin/bgp268.

15. Hausen zur H (1999) Immortalization of human cells and their malignant conversion by high risk human papillomavirus genotypes. Semin Cancer Biol 9: 405–411. doi:10.1006/scbi.1999.0144.

16. Duensing A, Spardy N, Chatterjee P, Zheng L, Parry J, et al. (2009) Centrosome overduplication, chromosomal instability, and human papillomavirus oncoproteins. Environ Mol Mutagen 50: 741–747. doi:10.1002/em.20478.

17. Duensing S, Lee LY, Duensing A, Basile J, Piboonniyom S, et al. (2000) The human papillomavirus type 16 E6 and E7 oncoproteins cooperate to induce mitotic defects and genomic instability by uncoupling centrosome duplication from the cell division cycle. Proc Natl Acad Sci USA 97: 10002–10007. doi:10.1073/pnas.170093297.

18. Patel D, Incassati A, Wang N, McCance DJ (2004) Human papillomavirus type 16 E6 and E7 cause polyploidy in human keratinocytes and up-regulation of G2-M-phase proteins. Cancer Res 64: 1299–1306.

19. Finzer P, Aguilar-Lemarroy A, Rösl F (2002) The role of human papillomavirus oncoproteins E6 and E7 in apoptosis. Cancer Lett 188: 15–24.

20. Duensing S, Münger K (2002) The human papillomavirus type 16 E6 and E7 oncoproteins independently induce numerical and structural chromosome instability. Cancer Res 62: 7075–7082.

21. Kessis TD, Connolly DC, Hedrick L, Cho KR (1996) Expression of HPV16 E6 or E7 increases integration of foreign DNA. - Abstract - UK PubMed Central. Oncogene.

22. Kadaja M, Isok-Paas H, Laos T, Ustav E, Ustav M (2009) Mechanism of genomic instability in cells infected with the high-risk human papillomaviruses. PLoS Pathog 5: e1000397. doi:10.1371/journal.ppat.1000397.

23. Eichten A, Rud DS, Grace M, Piboonniyom S-O, Zacny V, et al. (2004) Molecular pathways executing the “trophic sentinel” response in HPV-16 E7-expressing normal human diploid fibroblasts upon growth factor deprivation. Virology 319: 81–93. doi:10.1016/j.virol.2003.11.008.

24. Zehbe I, Richard C, DeCarlo CA, Shai A, Lambert PF, et al. (2009) Human papillomavirus 16 E6 variants differ in their dysregulation of human keratinocyte differentiation and apoptosis. Virology 383: 69–77. doi:10.1016/j.virol.2008.09.036.

25. Chang YE, Laimins LA (2000) Microarray analysis identifies interferon-inducible genes and Stat-1 as major transcriptional targets of human papillomavirus type 31. J Virol 74: 4174–4182.

26. Hebner C, Beglin M, Laimins LA (2007) Human papillomavirus E6 proteins mediate resistance to interferon-induced growth arrest through inhibition of p53 acetylation. J Virol 81: 12740–12747. doi:10.1128/JVI.00987-07.

27. Earle WR, Schilling EL, Stark TH, Straus NP, Brown MF, et al. (1943) Production of Malignancy in Vitro. IV. The Mouse Fibroblast Cultures and Changes Seen in the Living Cells. Journal of the ….

28. Gey GO (1954) Some aspects of the constitution and behavior of normal and malignant cells maintained in continuous culture. Harvey Lect 50: 154–229.

29. Gey GO, Coffman WD, Kubicek MT (1952) Tissue culture studies of the proliferative capacity of cervical carcinoma and normal epithelium. Cancer Res.

30. Culliton BJ (1974) HeLa Cells: Contaminating Cultures around the World. Science 184: 1058–1059. doi:10.1126/science.184.4141.1058.

31. Skloot R (2010) Immortal Life of Henrietta Lacks. MacMillan. pp.

32. Gartler SM (1967) Genetic markers as tracers in cell culture. Natl Cancer Inst Monogr 26: 167–195.

33. Gartler SM (1968) Apparent Hela cell contamination of human heteroploid cell lines. Nature 217: 750–751.

34. Nelson-Rees WA, Flandermeyer RR, Hawthorne PK (1974) Banded marker chromosomes as indicators of intraspecies cellular contamination. Science 184: 1093–1096.

35. Nelson-Rees WA, Flandermeyer RR (1976) HeLa cultures defined. Science 191: 96–98.

36. Nelson-Rees WA, Daniels DW, Flandermeyer RR (1981) Cross-contamination of cells in culture. Science 212: 446–452.

37. MacLeod RA, Dirks WG, Matsuo Y, Kaufmann M, Milch H, et al. (1999) Widespread intraspecies cross-contamination of human tumor cell lines arising at source. Int J Cancer 83: 555–563.

38. Lacroix M (2008) Persistent use of “false” cell lines. Int J Cancer 122: 1–4. doi:10.1002/ijc.23233.

39. Silva LM, Montes de Oca H, Diniz CR, Fortes-Dias CL (2001) Fingerprinting of cell lines by directed amplification of minisatellite-region DNA (DAMD). Braz J Med Biol Res 34: 1405–1410. doi:10.1590/S0100-879X2001001100005.

40. Masters JR, Thomson JA, Daly-Burns B, Reid YA, Dirks WG, et al. (2001) Short tandem repeat profiling provides an international reference standard for human cell lines. Proceedings of the ….

41. Capes-Davis A, Theodosopoulos G, Atkin I, Drexler HG, Kohara A, et al. (2010) Check your cultures! A list of cross-contaminated or misidentified cell lines. Int J Cancer 127: 1–8. doi:10.1002/ijc.25242.

42. Harris H, Watkins JF (1965) Hybrid Cells Derived From Mouse And Man: Artificial Heterokaryons Of Mammalian Cells From Different Species. Nature 205: 640–646.

43. Giannella RA, Washington O, Gemski P, Formal SB (1973) Invasion of HeLa cells by Salmonella typhimurium: a model for study of invasiveness of Salmonella. J Infect Dis 128: 69–75.

44. Lewinski N, Colvin V, Drezek R (2008) Cytotoxicity of nanoparticles. Small 4: 26–49. doi:10.1002/smll.200700595.

45. Villanueva A, Cañete M, Roca AG, Calero M, Veintemillas-Verdaguer S, et al. (2009) The influence of surface functionalization on the enhanced internalization of magnetic nanoparticles in cancer cells. Nanotechnology 20: 115103. doi:10.1088/0957-4484/20/11/115103.

46. Chaudhry MA, Chodosh LA, McKenna WG, Muschel RJ (2002) Gene expression profiling of HeLa cells in G1 or G2 phases. Oncogene 21: 1934–1942. doi:10.1038/sj.onc.1205264.

47. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, et al. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13: 1977–2000. doi:10.1091/mbc.02-02-0030.

48. Hnilicová J, Hozeifi S, Dušková E, Icha J, Tománková T, et al. (2011) Histone deacetylase activity modulates alternative splicing. PLoS ONE 6: e16727. doi:10.1371/journal.pone.0016727.

49. Murray JI, Whitfield ML, Trinklein ND, Myers RM, Brown PO, et al. (2004) Diverse and specific gene expression responses to stresses in cultured human cells. Mol Biol Cell 15: 2361–2374. doi:10.1091/mbc.E03-11-0799.

50. Ludwig H, Mages J, Staib C, Lehmann MH, Lang R, et al. (2005) Role of viral factor E3L in modified vaccinia virus ankara infection of human HeLa Cells: regulation of the virus life cycle and identification of differentially expressed host genes. J Virol 79: 2584–2596. doi:10.1128/JVI.79.4.2584-2596.2005.

51. Jaluria P, Betenbaugh M, Konstantopoulos K, Shiloach J (2007) Enhancement of cell proliferation in various mammalian cell lines by gene insertion of a cyclin-dependent kinase homolog. BMC Biotechnol 7: 71. doi:10.1186/1472-6750-7-71.

52. Alekseev OM, Richardson RT, Alekseev O, O'Rand MG (2009) Analysis of gene expression profiles in HeLa cells in response to overexpression or siRNA-mediated depletion of NASP. Reprod Biol Endocrinol 7: 45. doi:10.1186/1477-7827-7-45.

53. Kim H, Chen J, Yu X (2007) Ubiquitin-binding protein RAP80 mediates BRCA1-dependent DNA damage response. Science 316: 1202–1205. doi:10.1126/science.1139621.

54. Kittler R, Putz G, Pelletier L, Poser I, Heninger A-K, et al. (2004) An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. Nature 432: 1036–1040. doi:10.1038/nature03159.

55. Zhu C, Zhao J, Bibikova M, Leverson JD, Bossy-Wetzel E, et al. (2005) Functional analysis of human microtubule-based motor proteins, the kinesins and dyneins, in mitosis/cytokinesis using RNA interference. Mol Biol Cell 16: 3187–3199. doi:10.1091/mbc.E05-02-0167.

56. Pelkmans L, Fava E, Grabner H, Hannus M, Habermann B, et al. (2005) Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. Nature 436: 78–86. doi:10.1038/nature03571.

57. Neumann B, Walter T, Hériché J-K, Bulkescher J, Erfle H, et al. (2010) Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature 464: 721–727. doi:10.1038/nature08869.

58. Fuchs F, Pau G, Kranz D, Sklyar O, Budjan C, et al. (2010) Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. Mol Syst Biol 6: 370. doi:10.1038/msb.2010.25.

59. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945. doi:10.1038/nature03001.

60. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31–46. doi:10.1038/nrg2626.

61. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63. doi:10.1038/nrg2484.

62. Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nat Meth 5: 19–21. doi:10.1038/nmeth1157.

63. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. Clin Chem 55: 856–866. doi:10.1373/clinchem.2008.107565.

64. Wu M-Y, Eldin KW, Beaudet AL (2008) Identification of chromatin remodeling genes Arid4a and Arid4b as leukemia suppressor genes. J Natl Cancer Inst 100: 1247–1259. doi:10.1093/jnci/djn253.

65. Affymetrix ENCODE Transcriptome Project, Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009) Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature 457: 1028–1032. doi:10.1038/nature07759.

66. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, et al. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 7: 548. doi:10.1038/msb.2011.81.

67. Chen TR (1988) Re-evaluation of HeLa, HeLa S3, and HEp-2 karyotypes. Cytogenet Cell Genet 48: 19–24.

68. Francke U, Hammond DS, Schneider JA (1973) The band patterns of twelve D 98-AH-2 marker chromosomes and their use for identification of intraspecific cell hybrids. Chromosoma 41: 111–121.

69. Heneen WK (1976) HeLa cells and their possible contamination of other cell lines: karyotype studies. Hereditas 82: 217–248.

70. Kraemer PM, Deaven LL, Crissman HA, Steinkamp JA, Petersen DF (1974) On the nature of heteroploidy. Cold Spring Harb Symp Quant Biol 38: 133–144.

71. Nelson-Rees WA, Hunter L, Darlington GJ, O'Brien SJ (1980) Characteristics of HeLa strains: permanent vs. variable features. Cytogenet Cell Genet 27: 216–231.

72. Popescu NC, DiPaolo JA (1989) Preferential sites for viral integration on mammalian genome. Cancer Genet Cytogenet 42: 157–171.

73. Ruess D, Ye LZ, Grond-Ginsbach C (1993) HeLa D98/aH-2 studied by chromosome painting and conventional cytogenetical techniques. Chromosoma 102: 473–477.

74. Stanbridge EJ, Flandermeyer RR, Daniels DW, Nelson-Rees WA (1981) Specific chromosome loss associated with the expression of tumorigenicity in human cell hybrids. Somatic Cell Genet 7: 699–712.

75. Macville M, Schröck E, Padilla-Nash H, Keck C, Ghadimi BM, et al. (1999) Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. Cancer Res 59: 141–150.

76. Mincheva A, Gissmann L, Hausen zur H (1987) Chromosomal integration sites of human papillomavirus DNA in three cervical cancer cell lines mapped by in situ hybridization. Med Microbiol Immunol 176: 245–256.

77. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85–97. doi:10.1038/nrg1767.

78. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. Nat Genet 37: 727–732. doi:10.1038/ng1562.

79. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, et al. (2004)

Detection of large-scale variation in the human genome. Nat Genet 36: 949–951. doi:10.1038/ng1416.

80. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–454. doi:10.1038/nature05329.

81. Levy S, Strausberg RL (2008) Human genetics: Individual genomes diversify. Nature 456: 49–51. doi:10.1038/456049a.

82. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, et al. (2010) Towards a comprehensive structural variation map of an individual human genome. Genome Biol 11: R52. doi:10.1186/gb-2010-11-5-r52.

83. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470: 59–65. doi:10.1038/nature09708.

84. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, et al. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40: 722–729. doi:10.1038/ng.128.

85. Stephens PJ, McBride DJ, Lin M-L, Varela I, Pleasance ED, et al. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 462: 1005–1010. doi:10.1038/nature08645.

86. Ding L, Ellis MJ, Li S, Larson DE, Chen K, et al. (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 464: 999–1005. doi:10.1038/nature08989.

87. Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, et al. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463: 191–196. doi:10.1038/nature08658.

88. Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, et al. (2010) The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature 467: 1109–1113. doi:10.1038/nature09460.

89. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38: 75–81. doi:10.1038/ng1697.

90. McCarroll SA (2010) Copy number variation and human genome maps. Nat Genet 42: 365–366. doi:10.1038/ng0510-365.

91. Koike A, Nishida N, Yamashita D, Tokunaga K (2011) Comparative analysis of copy number variation detection methods and database construction. BMC genetics 12: 29. doi:10.1186/1471-2156-12-29.

92. McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, et al. (2008)

Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet 40: 1107–1112. doi:10.1038/ng.215.

93. Willer CJ, Speliotes EK, Loos RJF, Li S, Lindgren CM, et al. (2009) Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 41: 25–34. doi:10.1038/ng.287.

94. de Cid R, Riveira-Munoz E, Zeeuwen PLJM, Robarge J, Liao W, et al. (2009) Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet 41: 211–215. doi:10.1038/ng.313.

95. Lee C, Iafrate AJ, Brothman AR (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nat Genet 39: S48–S54. doi:10.1038/ng2092.

96. Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements. Pathogenetics 1: 4. doi:10.1186/1755-8417-1-4.

97. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 84: 148–161. doi:10.1016/j.ajhg.2008.12.014.

98. Greshock J, Feng B, Nogueira C, Ivanova E, Perna I, et al. (2007) A comparison of DNA copy number profiling platforms. Cancer Res 67: 10173–10180. doi:10.1158/0008-5472.CAN-07-2102.

99. Hehir-Kwa JY, Egmont-Petersen M, Janssen IM, Smeets D, van Kessel AG, et al. (2007) Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis. DNA Res 14: 1–11. doi:10.1093/dnares/dsm002.

100. Winchester L, Yau C, Ragoussis J (2009) Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic 8: 353–366. doi:10.1093/bfgp/elp017.

101. Dellinger AE, Saw S-M, Goh LK, Seielstad M, Young TL, et al. (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Research 38: e105. doi:10.1093/nar/gkq040.

102. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25: 2865–2871. doi:10.1093/bioinformatics/btp394.

103. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Meth 8: 61–65. doi:10.1038/nmeth.1527.

104. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363–376. doi:10.1038/nrg2958.

105. Crasta K, Ganem NJ, Dagher R, Lantermann AB, Ivanova EV, et al. (2012) DNA breaks and chromosome pulverization from errors in mitosis. Nature 482: 53–58. doi:10.1038/nature10802.

106. Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11: 685–696. doi:10.1038/nrg2841.

107. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933. doi:10.1038/35057149.

108. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351. doi:10.1126/science.1058040.

109. Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, et al. (2000) The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 66: 979–988. doi:10.1086/302825.

110. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21. doi:10.1038/456018a.

111. Chang B-L, Cramer SD, Wiklund F, Isaacs SD, Stevens VL, et al. (2009) Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. Hum Mol Genet 18: 1368–1375. doi:10.1093/hmg/ddp035.

112. Meyer KB, Maia A-T, O'Reilly M, Teschendorff AE, Chin S-F, et al. (2008) Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. PLoS Biol 6: e108. doi:10.1371/journal.pbio.0060108.

113. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073. doi:10.1038/nature09534.

114. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272–276. doi:10.1038/nature08250.

115. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, et al. (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106: 19096–19101. doi:10.1073/pnas.0910672106.

116. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636–639. doi:10.1126/science.1186802.

117. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7: 61–80. doi:10.1146/annurev.genom.7.080505.115630.

118. Steigemann P, Wurzenberger C, Schmitz MHA, Held M, Guizetti J, et al. (2009) Aurora B-mediated abscission checkpoint protects against tetraploidization. Cell 136: 473–484. doi:10.1016/j.cell.2008.12.020.

119. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59. doi:10.1038/nature07517.

120. Yoon OK, Brem RB (2010) Noncanonical transcript forms in yeast and their regulation during environmental stress. RNA 16: 1256–1267. doi:10.1261/rna.2038810.

121. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, et al. (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Research 37: e123. doi:10.1093/nar/gkp596.

122. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 26: 873–881. doi:10.1093/bioinformatics/btq057.

123. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657–663. doi:10.1093/bioinformatics/btl646.

124. McKenna A, Hanna M, Banks E, Sivachenko A (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.

125. The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796. doi:10.1038/nature02168.

126. Rausch T, Zichner T, Schlattl A, M SA, Benes V, et al. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28: i333–i339. doi:10.1093/bioinformatics/bts378.

127. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318: 420–426. doi:10.1126/science.1149504.

128. Jones DTW, Jäger N, Kool M, Zichner T, Hutter B, et al. (2012) Dissecting the genomic complexity underlying medulloblastoma. Nature 488: 100–105. doi:10.1038/nature11284.

129. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462–467. doi:10.1159/000084979.

130. Jentsch I, Geigl J, Klein CA, Speicher MR (2003) Seven-fluorochrome mouse M-FISH for high-resolution analysis of interchromosomal rearrangements. Cytogenet Genome Res 103: 84–88.

131. Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, et al. (2004) National center for biotechnology information viral genomes project. J Virol 78: 7291–7298. doi:10.1128/JVI.78.14.7291-7298.2004.

132. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106. doi:10.1186/gb-2010-11-10-r106.

133. Bauer S, Robinson PN, Gagneur J (2011) Model-based Gene Set Analysis for Bioconductor. Bioinformatics. doi:10.1093/bioinformatics/btr296.

134. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639–1645. doi:10.1101/gr.092759.109.

135. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, et al. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 144: 27–40. doi:10.1016/j.cell.2010.11.055.

136. Rausch T, Jones DTW, Zapatka M, M SA, Zichner T, et al. (2012) Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. Cell 148: 59–71. doi:10.1016/j.cell.2011.12.013.

137. Maher CA, Wilson RK (2012) Chromothripsis and human disease: piecing together the shattering process. Cell 148: 29–32. doi:10.1016/j.cell.2012.01.006.

138. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254. doi:10.1371/journal.pbio.0050254.

139. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872–876. doi:10.1038/nature06884.

140. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144: 646–674. doi:10.1016/j.cell.2011.02.013.

141. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research 31: 3812–3814.

142. Mitra AB, Murty VV, Li RG, Pratap M, Luthra UK, et al. (1994) Allelotype analysis of cervical carcinoma. Cancer Res 54: 4481–4487.

143. Mullokandov MR, Kholodilov NG, Atkin NB, Burk RD, Johnson AB, et al. (1996) Genomic alterations in cervical carcinoma: losses of chromosome heterozygosity and human papilloma virus tumor status. Cancer Res 56: 197–205.

144. Rader JS, Gerhard DS, O'Sullivan MJ, Li Y, Li L, et al. (1998) Cervical intraepithelial neoplasia III shows frequent allelic loss in 3p and 6p. Genes Chromosomes Cancer 22: 57–65.

145. Koopman LA, Corver WE, van der Slik AR, Giphart MJ, Fleuren GJ (2000) Multiple genetic alterations cause frequent and heterogeneous human histocompatibility leukocyte antigen class I loss in cervical cancer. J Exp Med 191: 961–976.

146. Vermeulen CFW, Jordanova ES, Zomerdijk-Nooijen YA, Haar ter NT, Peters AAW, et al. (2005) Frequent HLA class I loss is an early event in cervical carcinogenesis. Hum Immunol 66: 1167–1173. doi:10.1016/j.humimm.2005.10.011.

147. Corver WE, Haar Ter NT, Fleuren GJ, Oosting J (2011) Cervical carcinoma-associated fibroblasts are DNA diploid and do not show evidence for somatic genetic alterations. Cell Oncol (Dordr) 34: 553–563. doi:10.1007/s13402-011-0061-5.

148. Kaelbling M, Klinger HP (1986) Suppression of tumorigenicity in somatic cell hybrids. Cytogenet Genome Res 41: 65–70. doi:10.1159/000132206.

149. Saxon PJ, Srivatsan ES, Stanbridge EJ (1986) Introduction of human chromosome 11 via microcell transfer controls tumorigenic expression of HeLa cells. EMBO J 5: 3461–3466.

150. Srivatsan ES, Benedict WF, Stanbridge EJ (1986) Implication of chromosome 11 in the suppression of neoplastic expression in human cell hybrids. Cancer Res 46: 6174–6179.

151. Koi M, Morita H, Yamada H, Satoh H, Barrett JC, et al. (1989) Normal human chromosome 11 suppresses tumorigenicity of human cervical tumor cell line SiHa. Mol Carcinog 2: 12–21.

152. Oshimura M, Kugoh H, Koi M, Shimizu M, Yamada H, et al. (1990) Transfer of a normal human chromosome 11 suppresses tumorigenicity of some but not all tumor cell lines. J Cell Biochem 42: 135–142. doi:10.1002/jcb.240420304.

153. Srivatsan ES, Chakrabarti R, Zainabadi K, Pack SD, Benyamini P, et al. (2002) Localization of deletion to a 300 Kb interval of chromosome 11q13 in cervical cancer. Oncogene 21: 5631–5642. doi:10.1038/sj.onc.1205698.

154. Srivatsan ES, Ying KL, Seeger RC (1993) Deletion of chromosome 11 and of 14q sequences in neuroblastoma. Genes Chromosomes Cancer 7: 32–37.

155. Zhuang Z, Merino MJ, Chuaqui R, Liotta LA, Emmert-Buck MR (1995) Identical Allelic Loss on Chromosome 11q13 in Microdissected in Situ and Invasive Human Breast Cancer. Cancer Res.

156. Chakrabarti R, Srivatsan ES, Wood TF, Eubanks PJ, Ebrahimi SA, et al. (1998) Deletion mapping of endocrine tumors localizes a second tumor suppressor gene on chromosome band 11q13. Genes Chromosomes Cancer 22: 130–137.

157. Tanaka C, Kimura T, Yang P, Moritani M, Yamaoka T, et al. (1998) Analysis of loss of heterozygosity on chromosome 11 and infrequent inactivation of the MEN1 gene in sporadic pituitary adenomas. J Clin Endocrinol Metab 83: 2631–2634.

158. Venugopalan M, Wood TF, Wilczynski SP, Sen S, Peters J, et al. (1998) Loss of heterozygosity in squamous cell carcinomas of the head and neck defines a tumor suppressor gene region on 11q13. Cancer Genet Cytogenet 104: 124–132.

159. Cheng Y, Chakrabarti R, Garcia-Barcelo M, Ha TJ, Srivatsan ES, et al. (2002) Mapping of nasopharyngeal carcinoma tumor-suppressive activity to a 1.8-megabase region of chromosome band 11q13. Genes Chromosomes Cancer 34: 97–103.

160. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. doi:10.1038/75556.

161. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27–30.

162. Sørensen CS, Syljuåsen RG (2012) Safeguarding genome integrity: the checkpoint kinases ATR, CHK1 and WEE1 restrain CDK activity during normal DNA replication. Nucleic Acids Research 40: 477–486. doi:10.1093/nar/gkr697.

163. DePamphilis ML (2005) Cell cycle dependent regulation of the origin recognition complex. Cell Cycle 4: 70–79.

164. Tavera-Mendoza LE, Wang T-T, White JH (2006) p19INK4D and cell death. Cell Cycle 5: 596–598.

165. Polager S, Ginsberg D (2009) p53 and E2f: partners in life and death. Nat Rev Cancer 9: 738–748. doi:10.1038/nrc2718.

166. Morrison DK (2009) The 14-3-3 proteins: integrators of diverse signaling cues that impact cell fate and cancer development. Trends Cell Biol 19: 16–23. doi:10.1016/j.tcb.2008.10.003.

167. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, et al. (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. Hum Mutat 28: 622–629. doi:10.1002/humu.20495.

168. Jackson SP, Bartek J (2009) The DNA-damage response in human biology and disease. Nature 461: 1071–1078. doi:10.1038/nature08467.

169. Khanna KK, Jackson SP (2001) DNA double-strand breaks: signaling, repair and the cancer connection. Nat Genet.

170. Mazur S, Csucs G, Kozak K (2012) RNAiAtlas: a database for RNAi (siRNA) libraries and their specificity. Database (Oxford) 2012: bas027. doi:10.1093/database/bas027.

171. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29: 308–311.

172. Iskow RC, Gokcumen O, Lee C (2012) Exploring the role of copy number variants in human adaptation. Trends Genet 28: 245–257. doi:10.1016/j.tig.2012.03.002.

173. Schulze S, Petersen I (2011) Gender and ploidy in cancer survival. Cell Oncol (Dordr) 34: 199–208. doi:10.1007/s13402-011-0013-0.

174. de Villiers E-M, Fauquet C, Broker TR, Bernard H-U, Hausen zur H (2004) Classification of papillomaviruses. Virology 324: 17–27. doi:10.1016/j.virol.2004.03.033.

175. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. Nature 489: 101–108. doi:10.1038/nature11233.

176. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth 5: 621–628. doi:10.1038/nmeth.1226.

177. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, et al. (2011) The genomic complexity of primary human prostate cancer. Nature 470: 214–220. doi:10.1038/nature09744.

178. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, et al. (2011) Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet 43: 964–968. doi:10.1038/ng.936.

179. Magrangeas F, Avet-Loiseau H, Munshi NC, Minvielle S (2011) Chromothripsis identifies a rare and aggressive entity among newly diagnosed multiple myeloma patients. Blood 118: 675–678. doi:10.1182/blood-2011-03-344069.

180. Kloosterman WP, Hoogstraat M, Paling O, Tavakoli-Yaraki M, Renkens I, et al. (2011) Chromothripsis is a common mechanism driving genomic rearrangements in primary and metastatic colorectal cancer. Genome Biol 12: R103. doi:10.1186/gb-2011-12-10-r103.

181. Carmena M, Ruchaud S, Earnshaw WC (2009) Making the Auroras glow: regulation of Aurora A and B kinase function by interacting proteins. Curr Opin Cell Biol 21: 796–805. doi:10.1016/j.ceb.2009.09.008.

182. Archambault V, Glover DM (2009) Polo-like kinases: conservation and divergence in their functions and regulation. Nat Rev Mol Cell Biol 10: 265–275. doi:10.1038/nrm2653.

183. Bharadwaj R, Yu H (2004) The spindle checkpoint, aneuploidy, and cancer. Oncogene 23: 2016–2027. doi:10.1038/sj.onc.1207374.

184. Harper JW, Burton JL, Solomon MJ (2002) The anaphase-promoting complex: it's not just for mitosis any more. Genes Dev.

185. Lingle WL, Barrett SL, Negron VC, D'Assoro AB, Boeneman K, et al. (2002) Centrosome amplification drives chromosomal instability in breast tumor development. Proc Natl Acad Sci USA 99: 1978–1983. doi:10.1073/pnas.032479999.

186. Lengauer C, Kinzler KW, Vogelstein B (1997) Genetic instability in colorectal cancers. Nature 386: 623–627. doi:10.1038/386623a0.

187. Holland AJ, Cleveland DW (2009) Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nat Rev Mol Cell Biol 10: 478–487. doi:10.1038/nrm2718.

188. Deng X, Disteche CM (2010) Genomic responses to abnormal gene dosage: the X chromosome improved on a common strategy. PLoS Biol 8: e1000318. doi:10.1371/journal.pbio.1000318.

189. Aït Yahya-Graison E, Aubert J, Dauphinot L, Rivals I, Prieur M, et al. (2007) Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. Am J Hum Genet 81: 475–491. doi:10.1086/520000.

190. Zhang Y, Malone JH, Powell SK, Periwal V, Spana E, et al. (2010) Expression in aneuploid Drosophila S2 cells. PLoS Biol 8: e1000320. doi:10.1371/journal.pbio.1000320.

191. Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO (2011) Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome Res 21: 2004–2013. doi:10.1101/gr.122614.111.

192. Stingele S, Stoehr G, Peplowska K, Cox J, Mann M, et al. (2012) Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. Mol Syst Biol 8: 608. doi:10.1038/msb.2012.40.

193. Mohr S, Bakal C, Perrimon N (2010) Genomic screening with RNAi: results and challenges. Annual review of biochemistry.

194. Kittler R, Buchholz F (2005) Functional genomic analysis of cell division by endoribonuclease-prepared siRNAs. Cell Cycle 4: 564–567.

195. Kimura J, Nguyen ST, Liu H, Taira N, Miki Y, et al. (2008) A functional genome-wide RNAi screen identifies TAF1 as a regulator for apoptosis in response to genotoxic stress. Nucleic Acids Research 36: 5250–5259. doi:10.1093/nar/gkn506.

196. Krishnan MN, Ng A, Sukumaran B, Gilfoy FD, Uchil PD, et al. (2008) RNA interference screen for human genes associated with West Nile virus infection. Nature 455: 242–245. doi:10.1038/nature07207.

197. Tang W, Dodge M, Gundapaneni D, Michnoff C, Roth M, et al. (2008) A genome-wide RNAi screen for Wnt/beta-catenin pathway components identifies unexpected roles for TCF transcription factors in cancer. Proc Natl Acad Sci USA 105: 9697–9702. doi:10.1073/pnas.0804709105.

198. Zhou H, Xu M, Huang Q, Gates AT, Zhang XD, et al. (2008) Genome-scale RNAi screen for host factors required for HIV replication. Cell Host Microbe 4: 495–504. doi:10.1016/j.chom.2008.10.004.

199. Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, et al. (2008) Identification of host proteins required for HIV infection through a functional genomic screen. Science 319: 921–926. doi:10.1126/science.1152725.

200. Draviam VM, Stegmeier F, Nalepa G, Sowa ME, Chen J, et al. (2007) A functional genomic screen identifies a role for TAO1 kinase in spindle-checkpoint signalling. Nat Cell Biol 9: 556–564. doi:10.1038/ncb1569.

201. Sudbery I, Enright AJ, Fraser AG, Dunham I (2010) Systematic analysis of off-target effects in an RNAi screen reveals microRNAs affecting sensitivity to TRAIL-induced apoptosis. BMC Genomics 11: 175. doi:10.1186/1471-2164-11-175.

202. Simpson JC, Joggerst B, Laketa V, Verissimo F, Cetin C, et al. (2012) Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. Nat Cell Biol 14: 764–774. doi:10.1038/ncb2510.

203. Echeverri CJ, Perrimon N (2006) High-throughput RNAi screening in cultured cells: a user's guide. Nat Rev Genet 7: 373–384. doi:10.1038/nrg1836.

204. Boutros M, Ahringer J (2008) The art and design of genetic screens: RNA interference. Nat Rev Genet 9: 554–566. doi:10.1038/nrg2364.

205. Sigoillot FD, King RW (2011) Vigilance and validation: Keys to success in RNAi screening. ACS Chem Biol 6: 47–60. doi:10.1021/cb100358f.

206. Sioud M (2011) Promises and challenges in developing RNAi as a research tool and therapy. Methods Mol Biol 703: 173–187. doi:10.1007/978-1-59745-248-9_12.

207. Masters JR (2000) Human cancer cell lines: fact and fantasy. Nat Rev Mol Cell Biol 1: 233–236. doi:10.1038/35043102.

208. Green ED, Guyer MS, National Human Genome Research Institute (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470: 204–213. doi:10.1038/nature09764.

209. Lowry WE, Richter L, Yachechko R, Pyle AD, Tchieu J, et al. (2008) Generation of human induced pluripotent stem cells from dermal fibroblasts. Proc Natl Acad Sci USA 105: 2883–2888. doi:10.1073/pnas.0711983105.

210. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, et al. (2007) Induced pluripotent stem cell lines derived from human somatic cells. Science 318: 1917–1920. doi:10.1126/science.1151526.

211. Park I-H, Zhao R, West JA, Yabuuchi A, Huo H, et al. (2008) Reprogramming of human somatic cells to pluripotency with defined factors. Nature 451: 141–146. doi:10.1038/nature06534.

212. Blelloch R, Venere M, Yen J, Ramalho-Santos M (2007) Generation of induced pluripotent stem cells in the absence of drug selection. Cell Stem Cell 1: 245–247. doi:10.1016/j.stem.2007.08.008.

213. Takahashi K, Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126: 663–676. doi:10.1016/j.cell.2006.07.024.

214. Nishikawa S-I, Goldstein RA, Nierras CR (2008) The promise of human induced pluripotent stem cells for research and therapy. Nat Rev Mol Cell Biol 9: 725–729. doi:10.1038/nrm2466.

215. Holden C, Vogel G (2008) Cell biology. A seismic shift for stem cell research. Science 319: 560–563. doi:10.1126/science.319.5863.560.

216. Hanna J, Saha K, Pando B, van Zon J, Lengner C, et al. (2009) Direct cell reprogramming is a stochastic process amenable to acceleration. Nature. doi:10.1038/nature08592.

217. Simpson JT, Durbin R (2010) Efficient construction of an assembly string graph using the FM-index. Bioinformatics 26: i367–i373. doi:10.1093/bioinformatics/btq217.

218. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829. doi:10.1101/gr.074492.107.

219. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108: 1513–1518. doi:10.1073/pnas.1017351108.

220. Li R, Li Y, Zheng H, Luo R, Zhu H, et al. (2010) Building the sequence map of the human pan-genome. Nat Biotechnol 28: 57–63. doi:10.1038/nbt.1596.

221. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. doi:10.1038/ng.1028.

URLs:

[1] Ferlay J, Shin HR, Bray F, Forman D, Mathers C and Parkin DM. GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10. Lyon, France: International Agency for Research on Cancer; 2010. Available from: http://globocan.iarc.fr, accessed on 15/09/2012.

[2] DePristo M, and the Genome Sequencing and Analysis Group
The Broad Institute
http://www.broadinstitute.org/gatk/guide/topic?name=best-practices

[3] Accession numbers
http://steinmetzlab.embl.de/helaseq/
Username: helaseq
Password: helaseq