Comprehensive characterization of tandem repeat instability in 62 colorectal tumors

Comprehensive characterization of tandem repeat instability in 62 colorectal tumors

Koletou Malamati

Master thesis to partially fulfill the requirements of:

Master of Science (MSc) in Computational Biology & Bioinformatics (CBB)
Eidgenössische Technische Hochschule Zürich (ETH)
Universität Zürich (UZH) *(Joint Degree)*

December, 2014

Research performed at:

Institute of Evolutionary Biology and Environmental Studies, University of Zürich

Professor: Prof. Dr. Andreas Wagner

Supervisor: Dr. Tugce Bilgin Sonay

INDEX

1. Aknowledgements

I would like to express my gratitude to everyone who supported me throughout the course of this Master Thesis project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

A special gratitude I give to Prof. Dr. Andreas Wagner and Dr. Tugce Bilgin Sonay for their support and guidance at the Institute of Evolutionary Biology and Environmental Studies at University of Zürich, whose contribution in stimulating suggestions and encouragement helped me to coordinate my project especially in writing this report.

I would also like to thank all the people who provided me with the facilities being required and conductive conditions for my project.

And finally, I would like to acknowledge with much appreciation the crucial role of the Greek Foundation Bodossaki for granting me a scholarship in order to pursue my graduate studies at the MSc program of Computational Biology and Bioinformatics at ETH/UZH.

Thank you,

Malamati Koletou

2. Abstract

Tandem repeat polymorphisms are major contributors to genomic variation and pathogenic phenotypes. Colorectal cancer is an outcome of the accumulation of various molecular alterations in the genome. Repeat instability in cancer is an important factor in the process of tumorigenesis but it needs to be better studied. In this analysis, I investigate tandem repeat instability in 62 genomes of colorectal tumors and their matched normal tissues for the exonic regions of 18,439 genes. I find an increased de novo formation and loss of repeats in the tumor genomes compared to their matched-normal pair. Furthermore, I observe an absence of repeat variability in the copy number of matched repeats in the genome pairs. Additionally, repeats with higher copy numbers and shorter unit lengths show increased variability. Finally, I observe that cancer genes are more enriched with tandem repeats than other genes. Overall, although this study proves that there is some increased repeat instability in exonic regions of colorectal cancer, however it shows that more investigation is needed to provide a more precice picture of a genomic signature of carcinogenesis.

3. Introduction

3.1. Tandem Repeats

In biology, studying the relationship between the genotype and the phenotype of an organism is crucial for a complete understanding of functions of biological elements and how their malfunctioning associates with disease (Durbin et al., 2010). In order to understand this relationship, many studies investigate variation of DNA sequences and its contribution to distinct phenotypes. Of the variations most studied are single nucleotide polymorphisms (SNPs) (Wang et al., 1998; Sachidanandam et al., 2001) and structural variation caused by copy number variation of bigger or smaller parts of the genome (Pinkel et al., 1998; Stranger et al., 2007). The former refers to variation at a single position in a DNA sequence that occurs in more than 1% of the population (Durbin et al., 2010). The latter corresponds to structural variation of the genome, where the number of copies of the whole or a part of the genome is changed. Copy number variation could account for as much as 13% of the human genome (Stankiewicz and Lupski, 2010). On the last two decades, a genomic element that was previously perceived as “junk DNA” (Ohno, 1972) received the spotlight of the scientific community and found to contribute considerably to genotypic variation (Payseur et al., 2011; O’Dushlaine et al., 2005; O’Dushlaine and Shields, 2008; Willems et al., 2014). This genomic element is tandem repeats (TRs). They are DNA sequences where one repeat unit is repeated in tandem. TRs are highly abundant in the human genome.It has been shown that more than 30 percent of coding regions contain a tandem repeat (Legendre et al., 2007), and 17 percent of the tandem repeats in the human genome are polymorphic (Gemayel et al., 2010). According to the size of their repeat units, TRs are divided in two main categories; microsatellites that consist of repeats with less than 9 nucleotides (nt) and minisatellites with unit length from 10 nt to 100 nt (Denoeud et al., 2003; Näslund et al., 2005). The name “satellite” has a historical origin and refers to the process of “density-gradient centrifugal separation of genomic DNA” where TRs were originally identified as ”satellite bands” (Kit, 1961).

It is challenging to find out whether a sequence is a tandem repeat or a biologically irrelevant structure that occurred by chance. Various algorithms (TRF, Mreps, Sputnik,

ATRHunter, iMEx, T-reks etc) have been developed for the identification of repeats. These algorithms use specified thresholds based on some repeat characteristics, such as the unit length of the repeat pattern, the repeat unit number, that is, how many times a unit is repeated, and the purity of the repeats (i.e. exact matching between repeat units). In projects studying TRs and their effects on phenotype, the most widely used tool is the Tandem Repeat Finder (TRF), (O'Dushlaine et al., 2005; Legendre et al., 2007; Payseur et al., 2011; Willems et al., 2014) which performs much better than other algorithms by detecting more perfect repeats with a consistency of performance due to its flexibility in the parameter settings and probabilistic modeling approach (Lim et al., 2013).

3.1.1. *Tandem Repeat Variations*

TRs are highly unstable elements with mutation rates varying from 10^{-3} to 10^{-7} per cell division and around 10^{-2} to 10^{-5} per generation, significantly higher than that of other parts of the genome and especially point mutations (Gemayel et al., 2010). The main source of polymorphisms in TRs is a structural variation that stems from repeat unit number alterations, more specifically the addition or deletion of one or more repeat unitscauses an expansion or a contraction in the repeat sequence, respectively. Several models have been proposed to describe the molecular mechanism that causes this variation and how it is affected by repeat characteristics (Gemayel et al., 2010). The two most accepted models so far are strand-slippage replication and recombination (Gemayel et al., 2010; Ellegren, 2004). Strand slippage basically takes place during DNA replication when there is a mispairing event or when a double-strand breaks in arepeat tract (Fan and Chu, 2007). In the former case, the DNA strand that is synthesized anew misaligns with the template strand, causing formation of a loop (Lovett, 2004). This can result in a repeat contraction, if the template strand loops out, or in an expansion, if the nascent strand loops out (Gemayel et al., 2010; Lovett, 2004). The other major model for repeat unit number variation involves the meiotic recombination process (Gemayel et al., 2010; Richard and Pâques, 2000). Meiotic recombination is the crossing over between chromosomes, and although this model can explain repeat instability found in the germline, it cannot account for the majority of tandem repeat polymorphisms during somatic mutations (Richard and Pâques, 2000).

3.1.2. *Repeat characteristics of variable repeats*

Mutation rates differ among repeats (Weber and Wong, 1993; Brinkmann et al., 1998). There have been efforts to determine which repeat characteristics contribute most to their variability (O'Dushlaine et al., 2005; Legendre et al., 2007; O'Dushlaine and Shields, 2008; Payseur et al., 2011; Willems et al., 2014). The most significant contributor is repeat unit number, which seems to increase repeat variability exponentially (Legendre et al., 2007, O'Dushlaine et al., 2005; O'Dushlaine and Shields, 2008; Payseur et al., 2011, Willems et al., 2014). Moreover, several studies indicate higher repeat variability for repeats with greater TRF score (O'Dushlaine and Shields, 2008, Legendre et al., 2007) and purity (O'Dushlaine and Shields, 2008, Willems et al., 2014). Repeat unit length is also a significant contributor to repeat variability, although there is no consensus on the direction of the effect. Some studies revealed higher repeat variability for shorter repeat units (O'Dushlaine and Shields, 2008; Ellegren, 2004; Willems et al., 2014), which can be explained by their high mutation rates (Chakraborty et al. 1997; Kelkar et al., 2008). However other studies found signs of the opposite effect (Payseur et al., 2011, O'Dushlaine et al., 2005). In coding regions, the most polymorphic repeats are trimeric and hexameric repeats (i.e. with unit length of 3 nt and 6 nt, respectively) (Willems et al., 2014, refs), probably due to higher selection pressure in exons to avoid frameshift mutations (Ellegren, 2004), which are more destructive than point mutations (Duval and Hamelin, 2002). Coding TRs are, indeed much less variable than noncoding TRS (Willems et al., 2014, Payseur et al., 2011) TRs in introns, and TRs in untranslated regions (Payseur et al., 2011). In a genomic survey of repeat variation between two human genomes, Payseur and colleagues (Payseur et al., 2011), documented a negative effect of the repeat sequence GC content in the variability of the repeat.

3.1.3. *Tandem Repeats Confer Functional Variability and Cause Disease*

A large amount of research exists on the consequences of repeat unit number variation. TRs have been associated with various traits related to organismal evolvability (Kashi and King, 2006; Gemayel et al., 2010). They can cause phase variation in prokaryotes.

For example in the bacterium *Neisseria gonorrhoeae*, in some members of the P.II gene family that encode for a membrane signal peptide, the variation of the CTCTT repeat in the 5' region of the gene confers an ON/OFF switching mechanism that allows the bacteria to survive in the host during an infection (Stern et al., 1986). TRs can also generate functional variability in eukaryotic microbes. The best studied case is that of the FLO1 gene in *Saccharomyces cerevisiae,* which encodes for a cell-surface adhesion protein. Repeat unit number variations in the intergenic regions of this gene allow the organism to rapidly adapt to environmental changes and adjust its flocculation (i.e. the adherence to each other) (Verstrepen et al., 2005). Other examples of repeat-related phenotypic variation can be found in the control mechanisms of the circadian rhythm of Drosophila (Sawyer et al., 1997), and also in major skull morphology changes in canine species (Fondon and Garner, 2004). Irrefutable examples of experimental data suggest that TRs play an integral role in mammalian morphological evolution (Fondon and Garner, 2004). Additionally, TRs are used in population genetic studies to create genetic fingerprints (Kayser and de Knijff, 2011) and lineage databases (Khan and Mittelman, 2013).

One of the main discoveries that drew the attention of researchers to TRs is their role in disease formation. There is strong evidence that tandem repeats are the cause behind some monogenic disorders, including several neurodegenerative diseases, the most common of those being Huntington disease, Fragile X Syndrome, and myotonic dystrophy (López Castel et al., 2010; Gemayel et al., 2010). Huntington disease is caused by a CAG repeat unit number expansion in exon 1 of the IT15 gene, and it alters the expression of a protein named huntingtin (The Huntington's Disease Collaborative Research Group, 1993). Fragile X syndrome is caused by a CGG repeat unit number expansion in the 5' untranslated region of the FMRI gene which results in its transcriptional silencing (Verkerk et al., 1991). Finally, both types of myotonic dystrophy, type 1 and type 2, are caused by repeat unit number expansions (Gemayel et al., 2010). Type 1 myotonic dystrophy is caused by an expansion of a CTG repeat in the 3' untranslated region of the protein serine-threonine kinase (Brook et al., 1992) and type 2 is caused by the expansion of a CCTG repeat in intron 1 of the zinc finder protein encoding gene ZNF9 (Liquori et al., 2001).

Repeat unit number variation is not only associated with monogenic disorders, but it has also been associated with the manifestation of different types of cancer, including breast (Wood et al., 2007; McIver et al., 2014), colorectal (Wood et al., 2007; Boland and Goel, 2010; Vilar and Gruber, 2010), endometrial and gastric adenomas (Woerner et al., 2003). For example, a smaller number of CAG repeat units in the first exon of the androgen receptor gene has been associated with higher risk of prostate cancer, and specifically with higher risk of "distant metastatic and fatal prostate cancer" (Giovannucci et al., 1997). Also, a tetra-nucleotide (TTTA) repeat unt number expansion in the forth intron of the CYP19 gene has been associated with breast cancer risk (Haiman et al., 2000).

3.2. Cancer

Cancer is at its basis a genetic disease (Vogelstein and Kinzler, 2004; Futreal et al., 2004; Kan et al., 2010; Cancer & Atlas, 2012) and a lot of progress has been made to identify which are the mechanisms and mutations that initiate and drive cancer (Giovannucci et al., 1997; Haiman et al., 2000; Woerner et al., 2003; Fresno Vara et al., 2004; Vogelstein and Kinzler, 2004; Harris and Levine, 2005; Wood et al., 2007; Cancer & Atlas, 2012). In contrast to other genetic disorders, though, cancer is not caused by a single gene defect. Instead, a combination or rather an accumulation of mutations leads to carcinogenesis. These mutations occur in three main classes of genes: oncogenes, tumor suppressor genes, and genes associated with genomic instability, which, when mutated, increase the genomic mutation rate and can promote tumorigenesis in this way(Vogelstein and Kinzler, 2004, Davies et al., 2002). The progression from a normal cell to a malignant tumor is a multistep and variant process, but six hallmarks have been proposed to describe it: (i) self-sufficiency in growth signalling, (ii) insensitivity to anti-growth signals, (iii) evasion of apoptosis, (iv) enabling of a limitless replicative potential, (v) induction and sustainment of angiogenesis, (vi) activation of metastasis and invasion of tissue (Hanahan and Weinberg, 2000).

In the race to detect, prevent or cure cancer, there is a growing interest in identifying those mutations that are responsible for tumorigenesis. The mutations that occur in oncogenes lead to their activation and result in cell proliferation and survival of tumorous

cells. The types of mutations that activate oncogenes are usually chromosomal translocations, gene amplifications, or point mutations that can increase the expression of a gene product (Vogelstein and Kinzler, 2004). One such example is the mutation of BRAF, where the substitution of only one amino acid residue by another leads to overactive gene product and aberrant growth (Davies et al., 2002). Tumor suppressor genes regulate cell birth, differentiation, and death, whose inactivation through various mutations can allow tumor formation (Knudson, 2002; Vogelstein and Kinzler, 2004). Such mutations can occur through deletions or insertions, epigenetic silencing or other changes in gene regulation (Vogelstein and Kinzler, 2004). Both these classes of genes operate in a similar fashion, driving tumor progression by increasing cell proliferation, inhibiting apoptosis and enhancing angiogenesis (Nowell, 2002). Mutations in genes affecting genomic stability promote tumorigenesis in a different way. This class of genes is responsible for correcting any alteration that can occur in the genome, from single base substitutions to chromosomal recombination events (Vogelstein and Kinzler, 2004; Nowell, 2002). Therefore, when these genes are inactivated, the genomic mutation rate increases.

Although most current cancer research is focused on identifying mutations in genes that can lead to cancer (Davies et al., 2002; Woerner et al., 2003; Wood et al., 2007; Madsen et al., 2008; Vilar and Gruber, 2010; McIver et al., 2014), some other studies focus on the level of pathways that are central in cancer manifestation, instead (Vogelstein and Kinzler, 2004; Kan et al., 2010; Fearon, 2011). There are five pathwaysthat appear to be dysregulated in all cancers; the p53, Wnt, MARK, mTOR and TGF beta (Knudson, 2002; Vogelstein and Kinzler, 2004; Fresno Vara et al., 2004; Logan and Nusse, 2004; Kan et al., 2010; Cancer & Atlas, 2012). Some of these pathways play a role in cell proliferation, gene transcription and cell migration, such as the Wnt, MAPK and mTOR pathways, while others become active in cell death and apoptosis, such as the p53 and TGF beta pathways (Fresno Vara et al., 2004; Logan and Nusse, 2004; Harris and Levine, 2005; Kan et al., 2010; Cancer & Atlas, 2012). These pathways are not always altered in the same way between different cancer types, but it has been suggested that all cancers will eventually be shown to contain some mutations that affects them (Knudson, 2002).

All the mutations mentioned above can happen both in the germline (hereditary predisposition to cancer) or in somatic cells (sporadic tumors). Hereditary mutations can occur in every type of cancer but they are characterized by an overall increased probability of forming tumors and may not be causing cancer themselves (Knudson, 2002). Additional somatic mutations are still necessary for carcinogenesis and it is usually a non-affected parental allele that needs to be mutated (Vogelstein and Kinzler, 2004). In case of somatic mutations, some mutations are needed to initiate tumorigenesis process, whereas others are usually required for tumor progression (Vogelstein and Kinzler, 2004).

3.3. Colorectal Cancer

Colorectal cancer is the third most common cancer worldwide and the second most common cause of cancer-related deaths in the US (Siegel et al., 2014; UK, 2014). Like in any cancer, the two most common gene expression changes that occur in colorectal tumorigenesis are the activation of oncogenes that contribute to the proliferation of tumor cells and the inactivation of tumor suppressor genes that lead to uncontrollable growth of the tumor (Hanahan and Weinberg, 2000). There are some key driver oncogenes and tumor-suppressor genes that have been identified in a considerable fraction of colorectal cancers - such as APC, KRAS and p53 (Vogelstein et al., 1988) - but in the last two decades a lot more genes that show alterations in smaller subsets of colorectal cancers other than alterations in oncogenes and tumor-suppressor genes have been found (Ionov et al., 1993; Futreal et al., 2004; Jass, 2007; Imai and Yamamoto, 2008).

In general, two distinct kinds of mutations have been associated with colorectal cancers, although it has been proposed that they are not necessarily mutually exclusive (Imai and Yamamoto, 2008): Chromosomal instability, which is found in 85% of colorectal cancers and microsatellite instability (MSI) which is detected in 15% of colorectal cancers. Three percent of colorectal cancers are hereditary and the rest are sporadic (Boland and Goel, 2010). MSI causes a large number of microsatellite mutations, i.e. copy number changes throughout the genome (Boland et al., 1998; Boland and Goel, 2010; Fearon, 2011). It arises from defects in genes that affect genome instability and more specifically from alterations in the DNA mismatch repair system (Duval and Hamelin, 2002; Vilar and

Gruber, 2010). The mismatch repair system is evolutionarily highly conserved and responsible for recognizing and correcting mismatches and deletions or insertion loops during DNA replication (Vilar and Gruber, 2010). A genetic or epigenetic inactivation of the mismatch repair function and specifically mutations in proteins MLH1, MSH2, MSH3, MSH6, and PMS2 render the system defective, with severe consequences in the accumulation of errors in DNA, which in turn leads to MSI (Vilar and Gruber, 2010; Fearon, 2011). Both hereditary and sporadic cases of colorectal cancers that are associated with mismatch repair deficiency are characterized by MSI (Di Pietro et al., 2005). Specifically, the MSI phenotype in hereditary non-polyposis colorectal cancer is a result of a mutated gene from the mismatch repair family (hMLH1, hMSH2, hMSH6, hPMS2), whereas in sporadic colorectal cancers the mismatch repair deficiency in most cases arises from silencing of the hMLH1 gene (Duval and Hamelin, 2002).

3.4. Microsatellite Instability in Colorectal Cancer

MSI has been associated also with other types of cancer, such as gastric and endometrial cancers, but it has been mostly studied for colorectal cancers (Duval and Hamelin, 2002). The mutations that are actually responsible for the formation of tumors usually occur in some specific genes (Woerner et al., 2003), but the overall increased mutated incident confers a signature of tumors with the MSI phenotype. MSI was first identified independently from more than one group in 1993 (Thibodeau et al., 1993; Peltomäki et al., 1993; Aaltonen et al., 1993; Ionov et al., 1993) and since then there has been a lot of effort to measure and identify it (Boland et al., 1998; Brinkmann et al., 1998; Denoeud et al., 2003; Perucho, 2003; Legendre et al., 2007; Imai and Yamamoto, 2008; Boland and Goel, 2010; Payseur et al., 2011; Willems et al., 2014). In the National Cancer Institute workshop in 1997 (Boland et al., 1998), a panel of mononucleotide and dinucleotide markers was established to identify MSI subtypes. Specifically, MSI-H (i.e. high) is diagnosed through instability in more than one marker-locus of the initial 5-marker panel or in more than 30 percent of an extended panel. The MSI-L (i.e. low) subtype is characterized by only one unstable marker of the 5-marker panel or <30% in the extended one. Finally, if there is no identified instability a tumor is characterized as MSS (i.e. stable). Although the occurrence of high instability (MSI-H) has been correlated with many clinical and pathological parameters (Boland et al., 1998; Fearon,

2011), the distinction between the MSI-L and MSS may remain in controversion until further investigation provides more evidence on the matter (Boland and Goel, 2010).

In order to investigate repeats as suitable genetic markers, one has to take into consideration the different mutational and selective forces that act on different genomic locations (Ellegren, 2004). Overall, the genomic location of mutations is crucial in the manifestation of cancer (Duval and Hamelin, 2002; Perucho, 2003; Woerner et al., 2003; Wood et al., 2007; Fearon, 2011; Payseur et al., 2011). There are many genes, for example, where mutations in the coding region lead to cancer (Duval and Hamelin, 2002). An early example of a coding repeat that was associated with instability in human colorectal cancer was located in the TGFβRII gene and found to cause inactivating frameshift alterations (Markowitz et al., 1995). Other more common cases include the silencing of the adenomatous polyposis coli gene's expression, or the inactivation of p53 tumor-suppressor gene (Fearon and Vogelstein, 1990). The search for target genes containing coding repeats has increased considerably, because mutations in coding regions are more likely to disrupt the function of a gene (Duval and Hamelin, 2002). In line with that observation, it has been shown that coding regions contain less variable repeats than other locations in the genome (Payseur et al., 2011; Willems et al., 2014). Repeats in exons appear to be even less abundant (Tóth et al., 2000). This could be explained by selection against frameshift mutations in the coding regions that inhibit the expansion or contraction of any repeat tracts other than trinucleotide repeats (Ellegren, 2004). However, it has been suggested that MSI associated mutations are more likely to promote the growth of MMR-deficient cells and lead to MSI carcinogenesis if they occur in coding microsatellites of MMR related genes (Woerner et al., 2003) In my Master dissertation, I therefore studied tandem repeats and their variation in the exons of 62 colorectal tumors and their matched normal genomes from the same patient. I identified significantly more genes with tandem repeat gains and losses in tumors. I also determined repeat characteristics of stable and unstable repeats.

4. Methods

4.1. Genome Sequence Analysis

I downloaded the whole genome sequences of patients with colorectal cancer from both the tumor and the matched normal genome (from blood samples) of the same individual through authorized access to the Cancer Genome Atlas Data Portal (Cancer & Atlas, 2012) using the CGHub browser (Wilks et al., 2014) The data was chosen based on a recent study (Cancer Genome Atlas Network, 2012) analysing a subset of 97 colorectal carcinomas with low-pass (3-5X coverage) whole-genome sequences (Illumina HiSeq 2000). From these 97 tumor and matched-normal genome pairs I considered only those 62 pairs that had more than 90 percent of their exons aligned(see Supplementary Table S1).

Then I generated the consensus sequences for the exonic regions of all genomes using SAMtools (Release 0.1.19, 15 March 2013, (Li et al., 2009), based on the reference human genome build hg18 (March 2006 human reference sequence, NCBI Build 36.1, Lander et al., 2001). In order to specify the exonic regions, I considered all transcript variants for each gene in hg18. Moreover, I excluded those exons that contained transcript variants on more than one chromosome (i.e. transposons), which left me with 198,142 exons in a total of 18,439 genes. Whenever a gene had multiple transcripts, I included all exonic regions from all transcripts into one super-transcript (Madsen et al., 2008).

4.2. Tandem Repeat Identification

In order to identify genomic tandem repeat tracts I used the program Tandem Repeat Finder (TRF Version 4.07b) (Benson, 1999). I applied TRF using the following parameters: match=2, mismatch=5, indel=5, and default matching and indel probabilities of 0.80 and 0.10, respectively. I set the minimum alignment score to 40, since a recent large scale study on the variation of short human TRs based on the data from the first phase of the 1000 Genome Project (Siva, 2008) proved that genome-wide thresholds for TRF scores above 28-34 are sufficient to ensure that false positive repeats are limited to

be less than 1% of all identified repeats (Willems et al., 2014). Later, I further filtered the TRF results so that the matching frequency of the repeat units in the sequence is above 90 percent (e.g., at least 18 nucleotides of a repeat unit of 20 nucleotides must match the most common repeat unit in the whole repeat sequence). I also excluded any repeats with unit length longer than 100 nucleotides, because repeat units longer than that are not polymorphic (O'Dushlaine and Shields, 2008; Willems et al., 2014).

4.3. Investigating Tandem Repeat Polymorphism

4.3.1. *Repeat sets*

Once I had identified repeats in the tumor and the matched-normal genomes of each individual, I identified *matched* repeats, i.e. repeats that exist in both members of a genome pair, whose repeat unit consensus sequence is identical. For this purpose, I allowed a positional variation of the repeats up to 50 nucleotides within a gene between pairs of genomes because of the substantial shifting that can be caused by indels in a population (Durbin et al., 2010). In a few cases of conflicts, when a repeat in the normal genome matched more than one non-tandem repeat in the tumor within the 50 nucleotide window, I kept only one matched pair, either the closest one or the longest one. I then separated matched repeats into two groups: non-variable repeats, which have the exact same repeat unit number in the tumor and normal genome, and variable repeats, which differ in their repeat unit number between the pairs of genomes. I also identified repeats that are found only in one member of a genome pair, that is, either only in the tumor or in the normal genome. I called these repeats *unique* repeats. Gains and losses characterize the de novo formation and loss of repeats, respectively, in the tumor genome compared to the matched-normal one.

4.3.2. *Repeat unit number variation*

As repeat unit number variation I considered the repeat unit number differences between matched repeats. Moreover, I only considered repeat unit number differences greater than or equal to one. This clarification is necessary because Tandem Repeat Finder gives decimal numbers for repeat unit number values and it is possible to get a repeat

unit number difference less than one (e.g. the difference between a repeat unit number of 3.1 and 2.6 is 0.5). For each matched repeat I computed repeat unit number variation as the difference in repeat unit number between the tumor genome and its matched-normal genome.

4.3.3. *Genomic Variation*

Once I obtained unit number variation data for each repeat, I calculated an overall repeat unit number variation score for each genome pair.

For each genome pair I computed its (genomic) repeat unit number variation score in three different ways:
(1) as the mean of the repeat unit number differences of all repeats identified in a pair of genomes
(2) as the number of genes with at least one repeat with repeat unit number variation
(3) and as the number of tandem repeats with unit number variation
All the above values were calculated only for the set of variable repeats (i.e. repeat unit number variation greater than or equal to one). However, I was also interested in restricting my analysis to two additional sets; (i) expanded repeats (i.e. higher unit number in the tumor genome), and (ii) contracted repeats (i.e. lower unit number in the tumor genome). I therefore generated three genomic variation scores for each set of repeats (variable repeats, expansions, contractions).Thus, in total I calculated nine values for each of the 62 genome pairs, resulting in nine arrays of size 62 for the tumor-normal genome pairs.

4.3.4. *Control sample of normal/normal genome pairs*

I was also interested in having a control sample of genome pairs that would correspond to random variability between pairs of normal human genomes. For this purpose, I matched 62 normal genomes with each other once, and thus created a sample of 1891 normal-normal pairs. That is, I took all unordered pairs from 62 samples. I repeated all the computations described above for this set of genome pairs with the exception that I could not distinguish between expanded and contracted repeats, as well as between

repeat gains and losses. This was the case because in the control group I am considering unordered pairs of normal/normal genomes. Therefore, for the control group of normal pairs I repeated the repeat variation analysis only for the set of variable repeats. Additionally, for the de novo repeat gain and loss analysis I considered only the set of unique repeats.

5. Results

5.1. Tumor/normal pairs have significantly more genes with tandem repeats than the control sample of normal/normal pairs

First, I asked how many repeats a tumor genome has on average and whether this number is different than the number of repeats in its matched normal genome . To this end, I identified the repeats in every tumor and normal genome in my data set. On average, I found 7533 (±433, i.e., with a standard deviation of 433) repeats in a tumor genome and 7585 (±416) repeats in a normal genome. These numbers were statistically indistinguishable (Wilcoxon Rank Sum (WRS) test (Mann and Whitney, 1947), P=0.547). Because some genes contain more than one repeat, I further asked, how many genes contain tandem repeats in a tumor genome and in its matched normal genome. Again I found these numbers statistically indistinguishable (WRS test, P=0.5175). More specifically, among all 18,439 genes that I was investigating, I found repeats on average in 5278 (±240) or 28.6% of the genes in the tumor, and 5309 (±230) or 28.8% of the genes in the normal genome. From this initial investigation I observed that all analysed genomes, tumor and normal, did had a similar distribution of identified repeats, and a similar distribution of genes with repeats.

My analysis is based on investigating repeat instability, which includes both repeat unit number variations, and repeat gains and losses between the tumor and matched normal genomes. In order to conduct this analysis, I had to compare the sample with tumor/normal pairs with the control sample of normal/normal pairs. Therefore, I identified sets of repeats found within each genome pair. These sets consist of both matched repeats and repeats (see Methods for the definitions) unique to one of the two genomes of the pair. Considering all identified repeats within a genome pair, matched and unique, I found a significant difference between the number of genes with repeats in tumor/normal pairs compared to the number of genes with repeats in normal/normal pairs (WRS test, $P<10^{-16}$, see Figure 1). More specifically, on average 6248 (±158) or 33.9 (±0.9) percent of genes contained repeats in tumor/normal pairs and 5194 (±174) or 28.2 (±0.9) percent genes contained tandem repeats in the normal/normal pairs. However, this significant difference did not persist, when I considered the total number of

repeats instead of the number of genes with repeats (WRS test, P = 0.09609). More specifically, I identified 9842 (±318) repeats in tumor/normal genome pairs and 9807 (±480) repeats in normal/normal genome pairs. I will discuss possible reasons for this change in section six.

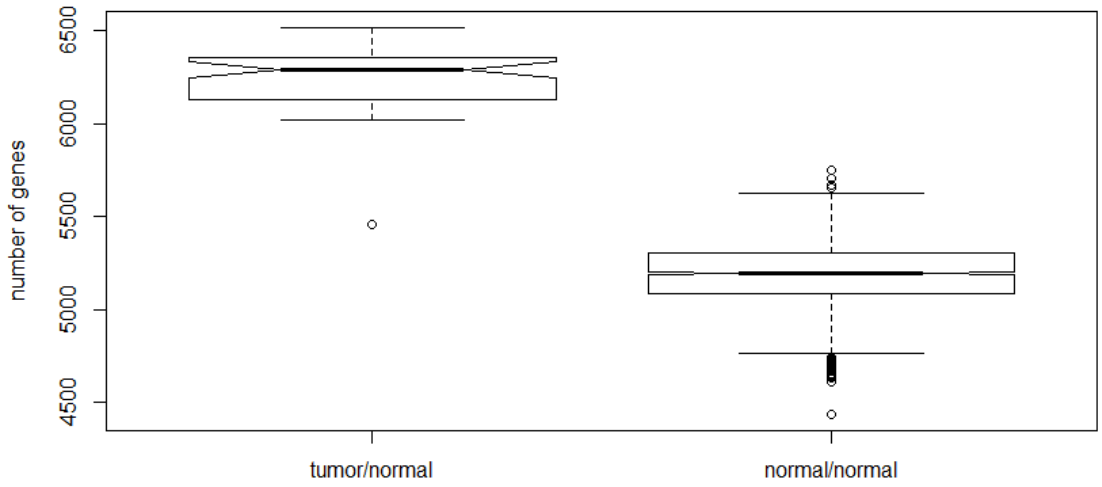

Figure 1. Tumor/normal pairs have significantly more genes with tandem repeats. Box plot of mean number of genes with identified tandem repeats in the tumor/normal pairs (mean=6248, n=62, left box) and in the normal/normal genome pairs (mean=5194, n=1891, right box). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range.

5.2. Tumor genomes underwent significantly more repeat gain and losses

To understand why I observed a difference between the number of genes with repeats in tumor/normal pairs and normal/normal genome pairs, I investigated cases of de novo repeat gain and repeat loss between genome pairs. To this end, I first identified matched repeats between all genome pairs (i.e. with the exact same repeat pattern, see Methods for details). I found that on average a tumor/normal genome pair contains 5275 (±544) matched repeats in 3992 (±332) genes and a normal/normal genome pair has on average 5188 (±438) matched repeats in 3946 (±270) genes. Hence, 53.5 (±4.8) percent of the repeats identified in the tumor/normal pairs and 52.9 (±3.4) percent of the repeats

identified in the normal/normal pairs could be matched to each other. The rest corresponded to either de novo repeat gains or losses. On average, there were 2258 (±302) gains in 1947 (±239) genes, and 2310 (±272) losses in 1981 (±210) genes, that is, a tumor/normal genome pair contained on average 4567 (±449) unique repeats, a number statistically indistinguishable (WRS test, P = 0.3458) from the number of unique repeats (4619±378) found in normal/normal genomes. When I asked, however how many genes contain unique repeats, I found a slight but significant enrichment (WRS test, P=0.046, see Figure 2) in tumor/normal pairs (3358±290 genes) compared to normal/normal pairs (3295±211).

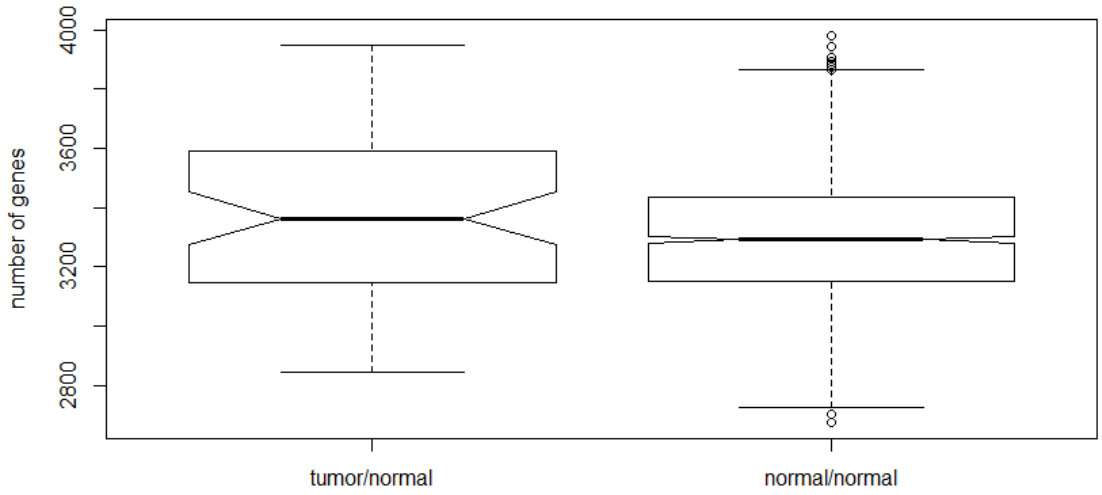

Figure 2. Tumor genomes underwent significantly more repeat gain and losses. Box plot of mean number of genes with unique tandem repeats in the tumor/normal pairs (mean=4567, n=62, left box) and in the normal/normal genome pairs (mean=4619, n=1891, right box). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range.

5.3. Most of the repeats have the same copy number between tumor and normal genomes

For the matched repeats I investigated their repeat unit number differences between tumor and matched normal genomes. I observed that variable repeats were significantly

less frequent than non-variable ones (WRS test, $P<10^{-16}$). Of the 5275 (±544) matched repeats, 185 (±23) that is, 3.6 (±0.8) percent were variable, whereas 5089 (±557) repeats were non-variable (see Figure 3). Furthermore, the percentage of repeat expansions (49.6±5.1) and contractions (50.4±5.1) within the variable repeats showed no statistical difference (WRS test, P = 0.2134).

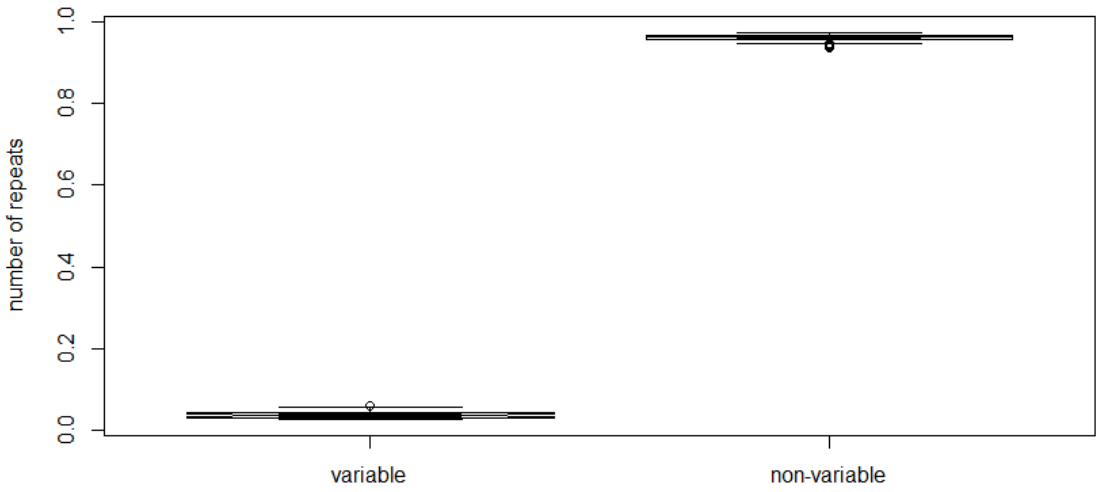

Figure 3. Most repeats have the same copy number between tumor and normal genomes. Box plot of mean number of repeats that have a different copy number between tumor and normal genomes (mean=185, n=5275, left box) and repeats that have the same copy number (mean=5089, n=5275, right box). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range.

Next, I wanted to estimate whether variable repeats were differentially represented between tumor/normal and the normal/normal pairs. For this purpose, I compared repeat unit number variation within the genome pairs but found no significant difference: both in tumor/normal the normal/normal pairs variable repeats varied by approximately 4.5 (±4.4) repeat units in their repat unit number.

5.4. Repeats are more frequent in cancer genes than in the total of 18,439 genes

So far, I had focused on tandem repeat instabilities in all genes. However, only a small minority of these unstable repeats may play a role in carcinogenesis. Many studies (Futreal et al., 2004; Vogelstein and Kinzler, 2004; Jass, 2007; Kan et al., 2010) identified genes that are likely to be relevant for cancer formation, such as genes that function in cell survival and cell death. In a next analysis, I therefore decided to study from my list of genes only a subset that is known to be cancer-related, amassed by Bilgin Sonay and colleagues (Bilgin Sonay et al., 2015). This set of genes consists of 371 genes (see Supplementary Table S2). Initially, all analysis between tumor/normal genomes and the normal/normal pairs produced similar results for the subset of cancer genesas it did for the total set of 18,439 genes. However, when I compared the number of genes with tandem repeats, both matched and unique, in the tumor/normal pairs, I found a significant enrichment in the cancer gene set, compared to the total set of genes (WRS test, $P<10^{-10}$). More specifically, on average 162 (±5) genes, that is, 43.6 (±1.5) percent of the cancer genes contained tandem repeats, whereas only 33.9 (±0.9) percent of all genes contained tandem repeats (see Figure 4a). Moreover, on average 6 (±2) genes, that is, 1.6 (±0.1) percent of cancer genes contained variable repeats, whereas only 1.0 (±0.1) percent of all genes contained variable repeats (see Figure 4b). Finally, on average 93 (±10) genes, that is 25.1 (±2.7) percent of the cancer genes contained unique repeats, whereas only 18.2 (±1.6) percent of all genes contained unique repeats (see Figure 4c).

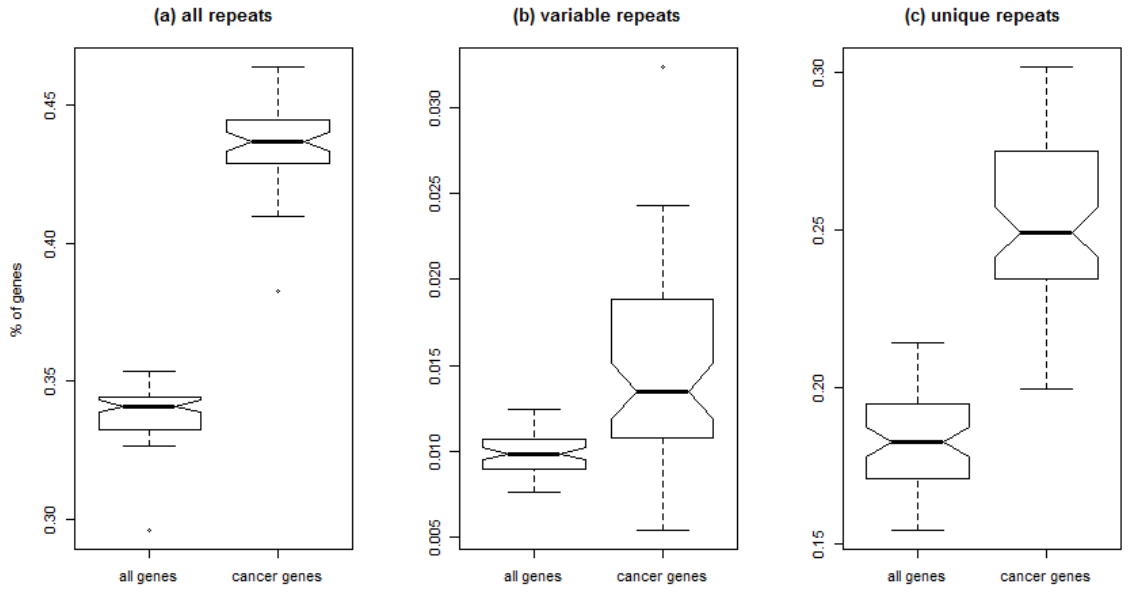

Figure 4. Repeats are more frequent in cancer genes than in the total of 18,439 genes. Each panel of box plots shows a statistical difference between the percentage of genes with repeats among the set of all 18,439 genes (left box in each panel) and the subset of 371 genes (right box in each panel); (a) genes with identified repeats (mean=33.9% in left box, mean=43.6% in right box), (b) genes with variable repeats (mean=1.0% in left box, mean=1.6% in right box) and (c) genes with unique repeats (mean=18.2% in left box, mean=25.1% in right box). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range.

5.5. Repeat Characteristics

From all 610,217 repeats identified in all genomes, I analysed the distributions of some important repeat characteristics and searched for any significant trends. The characteristics I considered were the repeat unit length, the repeat unit number, the TRF score, and the GC content of the repeat. Over all the identified repeats, I observed a wide variation of values for all these characteristics. Repeat unit length varied from 1 to 100 (which was an upper threshold I imposed, see Methods) with a mean of 11 (±10) nucleotides. The repeat unit number of repeats was between 2 and 55, with a mean of 6 (±7) units. The TRF score varied from 40, which was also a threshold I imposed (see Methods), to 1800 with mean of 61 (±59). Finally, the GC content ranged from 0 to 100

with a mean of 43 (±28) percent. Due to the significant differences I had observed in the sets of variant, non-variant, matched and unique repeats before, I decided to check the distribution of the characteristics in those sets separately.

5.5.1. *Variable repeats have significantly higher repeat unit number*

Repeat unit number is the most important contributor to repeat unit number variation (Legendre et al., 2007, O'Dushlaine et al., 2005; O'Dushlaine and Shields, 2008; Payseur et al., 2011, Willems et al., 2014). I computed the repeat unit number of variable (17.8±8.2), non-variable (6.6±7.9), matched (7.0±8.2), and unique repeats (5.2±6.1). The only significant difference I found was in the variable repeats. Their repeat unit number was significantly greater than the repeat unit number of all three other sets of repeats (WRS test, $P<10^{-16}$ for each of the comparisons between the set of variable repeats and each of the other sets, see Figure 5).

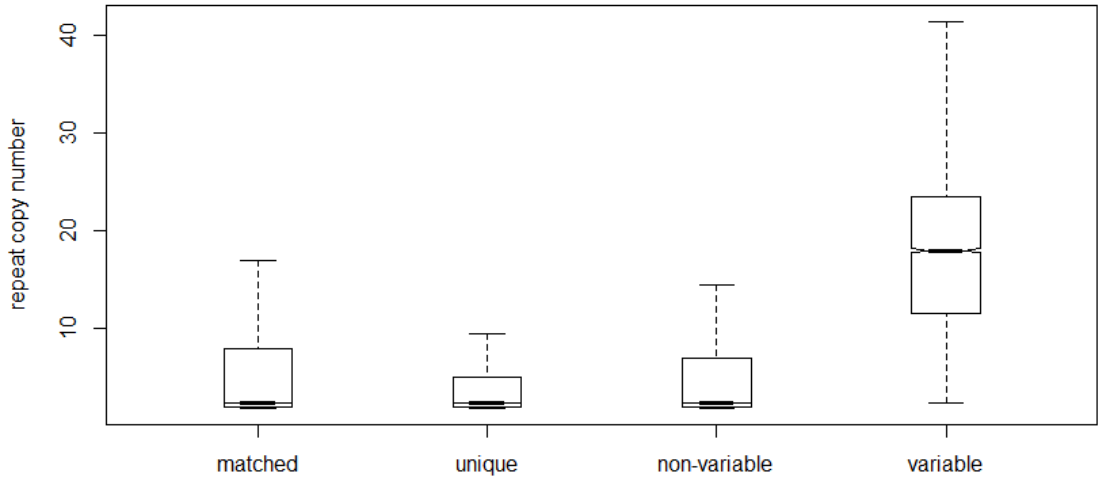

Figure 5. Variable repeats have significantly higher repeat unit number. Box plot of repeat unit number (from left to right boxes) for matched (mean=7.0, n=327,044), unique (mean=5.2, n=283,173), non-variable (mean=6.6, n=315,547) and variable repeats (mean=17.8, n=11,497). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range.

5.5.2. *Variable repeats have significantly greater TRF scores*

Many studies show that TRF scores can explain a great proportion of repeat unit number variation (O'Dushlaine and Shields, 2008, Legendre et al., 2007). As a further analysis, I therefore computed the TRF scores of the variable (67.1 ±53.7), non-variable (55.6±39.8), matched (55.1±39.1) and unique repeats (67.7±75.0). Variable repeats were the only set of repeats that had significantly greater TRF scores (WRS test, $P<10^{-16}$ for each of the comparisons between the set of variable repeats and each of the other sets, see Figure 6).

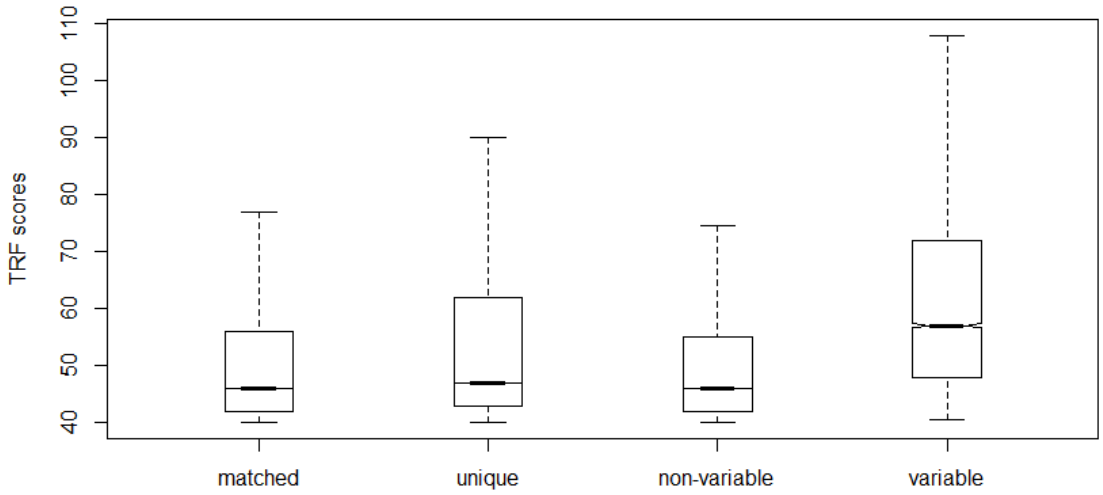

Figure 6. Variable repeats have significantly greater TRF scores. Box plot of repeat TRF score (from left to right) for matched (mean=55.1, n=327,044), unique (mean=67.7, n=283,173), non-variable (mean=55.6, n=315,547) and variable repeats (mean=67.1, n=11,497). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range.

5.5.3. *Unique repeats are significantly enriched for GC content*

Next, I compared the GC content between the variable (33.4±27.7), non-variable (35.9±26.5), matched (35.8±26.5) and unique repeats (50.5±28.8). Interestingly, I found that in all cases it was significantly higher for unique repeats (WRS test, $P<10^{-16}$ for each of the comparisons between the set of unique repeats and each of the other sets, see Figure 7).

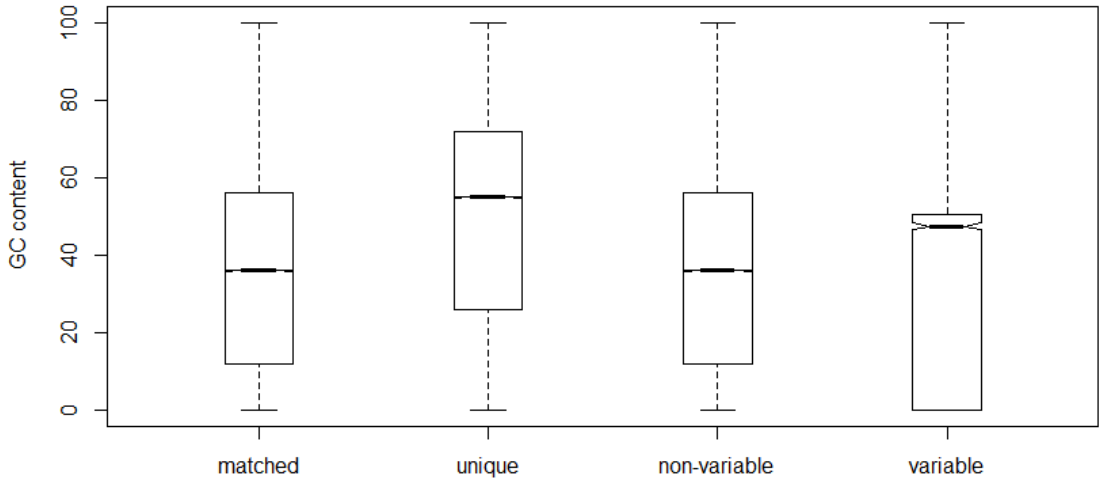

Figure 7. Unique repeats are significantly enriched for GC content. Box plot of repeat GC content (from left to right) for matched (mean=35.8, n=327,044), unique (mean=50.5, n=283,173), non-variable (mean=35.9, n=315,547) and variable repeats (mean=33.4, n=11,497). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range.

5.5.4. *Variable repeats have significantly smaller repeat units*

The direction in which the length of repeat units affects repeat unit number variation is still debated. Although multiple papers demonstrated greater variability for repeats with smaller repeat units (O'Dushlaine and Shields, 2008; Ellegren, 2004; Willems et al., 2014), a minority showed the opposite trend (Payseur et al., 2011, O'Dushlaine et al., 2005). I therefore decided to study repeat unit lengths in even greater detail. I discovered that repeats with units shorter than 7 nucleotides account for 94.7 percent of variable repeats, whereas the percentage of them is only 32 percent in the total number of identified repeats. To further illustrate that variable repeat have relatively shorter unit lenghts that the rest of the repeat set, I computed the unit length of the variable (mean: 3.1±5.4), non-variable (10.1±8.4), matched (9.9±8.4) and unique repeats (12.8±12.2). Variable repeats were significantly shorter than all three other sets of repeats (WRS test, $P<10^{-16}$ for each of the comparisons between the set of unique repeats and each of the other sets, see Figure 8) as I was expecting.

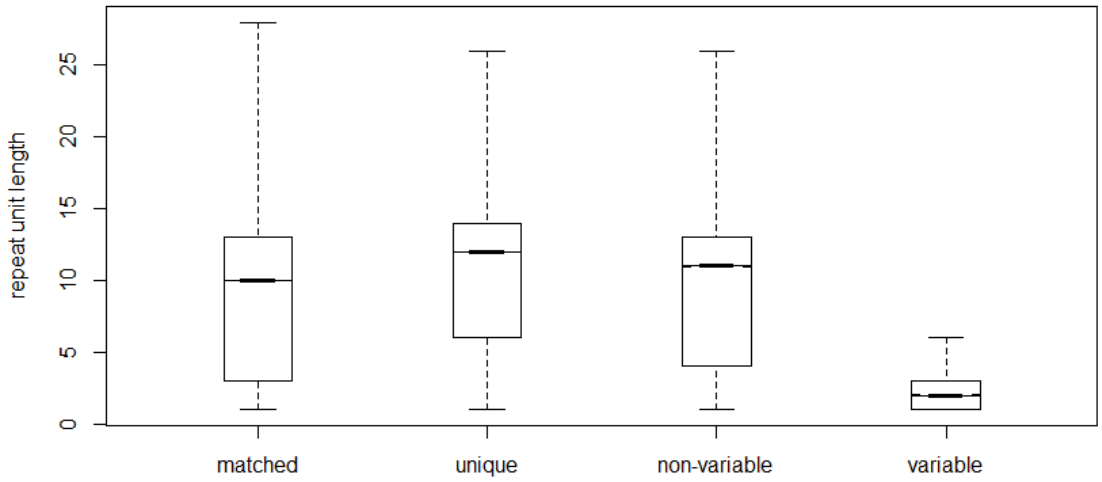

Figure 8. Variable repeats have significantly smaller repeat units. Box plot of the repeat unit length of the (from left to right) matched (mean=9.9, n=327,044), unique (mean=12.8, n=283,173), non-variable (mean=10.1, n=315,547) and variable repeats (mean=3.1, n=11,497). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3 percent of the data's range.

Multiple studies indicate biases for specific repeat unit lengths, i.e. repeat units comprising multiples of three nucleotides are found more frequently in exonic regions (Duval and Hamelin, 2002; Perucho, 2003; Woerner et al., 2003; Wood et al., 2007; Fearon, 2011; Payseur et al., 2011). Since I was investigating exonic regions, I wanted to check if I observe these biases in my data set as well. For this purpose I calculated the frequencies of all repeats for different repeat unit lengths and found that repeats that were multiples of three nucleotides account for 42.6 percent of all the repeats.

6. Discussion

I identified tandem repeats in genomes of colorectal tumor and healthy tissues of 62 patients in exonic regions of 18,439 human genes. In every genome, repeats identified were similar in frequency and the same observation occured for the frequency of genes with identified repeats in all genomes. Between a tumor/normal genome pair, I found on average 5275 matched and 4567 unique repeats. Genes with repeats, matched or unique, within a genome pair were significantly more abundant between tumor and matched normal genomes than between normal genome pairs. This was supported by the observation that genes with unique repeats were significantly abundant in the tumor/normal pairs. However I found no evidence for enrichment in the level of number of repeats in those genome pairs. This observation suggests that repeat gain and loss incidences might not vary within the genes that already contain repeats in the normal genomes but that they are acting on a larger set of genes in the tumor genomes. This might be due to the increased positive selection in tumor genomes compared to normal genomes. Hypermutability observed in the tumor genomes (Boland and Goel, 2010) supports this argument.

Although I observed some repeat unit number variation between the tumor and matched normal genomes, these cases were relatively fewer in the total number of identified repeats. Also, this variation occured without any preference for repeat expansions or contractions. In their study of variable repeats in human disease genes, Madsen and colleagues (Madsen et al., 2008) also do not report any preference for repeat expansions or contractions, although some other studies suggest a preference for expansions in genetic diseases (Tóth et al., 2000; López Castel et al., 2010; Gemayel et al., 2010). When I compared the incidences of repeat unit number variations between the tumor/normal genome pairs and the control sample, I found no enrichment in tumor genomes, which is surprising considering the increased number of repeat gain and loss events, along with the high mutability in most of these tumors. One possible explanation for this could be due to my conservative approach (see Methods) I underestimated cases of matched repeats and put them into the unique repeat category, which could decrease the sample size of variable repeats. Also, I was aligning genomes in my data set to the reference human genome to generate consensus sequences, which could

underestimate repat unit number variation of the identified repeats. This is a likely scenario if there are indels inside repeat sequence that lead to ambiguities in the consensus pattern of short repeat units (Madsen et al., 2007; Madsen et al., 2008). Also, an important limitation of this study comes from using low coverage genomes, which probably increases the false negative rate in the repeat detection.

Thanks to the recent technological advances in whole-genome sequencing (Zhao and Grant, 2011), searches for mutations in all genomic regions have become possible. A study about the genomic landscape of human breast and colorectal cancers by Wood and colleagues (Wood et al., 2007) suggests that most cancer related mutations, and especially the ones that are likely to be drivers, are not found in protein-coding genes but in non-coding regions. They support this argument on another paper by Beerenwinkel and colleagues claiming that tumor progression is promoted by a large number of mutations with small fitness advantage (Beerenwinkel et al., 2007). Eventually, studies like this end up with a large number of potential driver mutations but all these mutations may lie in only a few modified pathways (Vogelstein and Kinzler, 2004; Wood et al., 2007).

Besides searching for alterations in repeat tracts of coding regions in all known genes of the human genome, one can focus only on genes that are known to contribute to carcinogenesis. However, there is not a common consensus on what these genes are and evidence in the literature can be quite heterogeneous depending on tumor type (Vogelstein and Kinzler, 2004; Jass, 2007; Kan et al., 2010). For this reason, I used a set of cancer genes proposed by Bilgin Sonay and colleagues (Bilgin Sonay et al., 2015) that included genes associated with common pathway dysregulations appearing in most cancer types. I examined this collection of the cancer genes for repeat instabilities. I did not find any contradicting results to the previous analysis when I included all 18,439 genes. However, I observed an overall enrichment in genes with tandem repeats in general, and specifically in genes with variable and unique repeats, both in tumor and normal genomes. This enrichment may be due to the low coverage of the available genome sequences. Genes with poorly aligned sequences can lower the actual number of genes with repeats for the whole gene set and this could bias our analysis towards possibly better sequenced cancer genes which may appear to have more identified

repeats. However, it would require almost 2000 genes to be poorly aligned for this difference to be biased, which is highly unlikely considering the data I was analysing. Therefore, if this enrivhment in cancer genes is indeed an existing trend, it would be in agreement with another study conducted by Madsen et al. (Madsen et al., 2008). In that study, they showed for some diseases, including cancer, that disease-related genes have a significantly higher content of short tandem repeats. They suggested that this incidence could be indicative of a pathogenic phenotype and it be used for screening to detect rare mutations.

There has been a lot of research on repeat characteristics, centering on the question what makes a repeat more variable (O'Dushlaine et al., 2005; Legendre et al., 2007; O'Dushlaine and Shields, 2008; Payseur et al., 2011; Willems et al., 2014). Many patterns have been revealed, some of them contradictory. However what is common to these efforts is that they were all conducted on genomes from healthy tissues. Therefore, I was interested to investigate some of the most basic characteristics of repeats, in the cancer genomes analysed to check if they are in concordance with existing literature. This would be one of the rare analyses conducted on cancerous genomes (Woerner et al., 2003; Legendre et al., 2007; Madsen et al., 2008). As a final analysis, I therefore examined repeat characteristics for repeats with different instability status. The most irrefutable finding exiisting literature is that repeat unit number is significantly higher for variable repeats (O'Dushlaine et al., 2005; Legendre et al., 2007; O'Dushlaine and Shields, 2008; Payseur et al., 2011; Willems et al., 2014). This observation can be explained by the replication slippage mechanism that is more likely to add or remove a repeat unit in microsatellites with higher number of repeat units (Ellegren, 2004). In my analysis, it was also evident that the variable repeats had much higher repeat unit numbers than the rest. Another observation I made was that shorter repeat units were more frequent in variable repeats. Almost 95% of variable repeats had repeat units shorter than 7 nucleotides. The fact that shorter repeat units are more variable is also supported by most literature (O'Dushlaine and Shields, 2008; Kelkar et al., 2008; Willems et al., 2014). This could be because they are usually less stable and they have higher mutation rates (Chakraborty et al. 1997; Kelkar et al., 2008). I also showed that unique repeats had a significantly different GC content than other repeats, which could be the case because new sequence tracks introduced in the genome may

not have been affected by selection forces yet in order to adjust their GC content in a more favorable equilibrium as in the rest of the genome.
Furthermore, I found that almost half of the repeat unit lengths I identified were multiples of three nucleotides, as many earlier studies have documented (Duval and Hamelin, 2002; Perucho, 2003; Woerner et al., 2003; Wood et al., 2007; Fearon, 2011; Payseur et al., 2011). The likely reason is that repeats with unit lengths that are multiples of three are favored in exons because they do not introduce frameshift mutations, and therefore there are weak selective pressures against them (Duval and Hamelin, 2002; Ellegren, 2004).

My study is likely to be affected by the following limitations. First of all, the mutation mechanism of tandem repeats can be described by different models regarding whether it is causing somatic or germline mutations (Richard and Pâques, 2000; Lovett, 2004; Gemayel et al., 2010). Given the available data, it is not possible to distinguish between those types of mutations. Also, germline mutations in mismatch repair genes can cause colorectal cancer and increase the instance of subsequent somatic mutations that cause tumorigenesis progress (Vilar and Gruber, 2010; Fearon, 2011). Although, this is a distinction that I am not accounting for, somatic mutations are overall more frequent than the germline ones (Futreal et al., 2004). Therefore, I do not expect this to be a serious limitation. Moreover, the number of genomes I am analysing is limited to only 62, and in addition to that, genome alignment coverage could lead to underestimates of repeat unit numbers. This bias could decrease the incindence of repeat unit number variation in the genomes I was analysing. To detect repeat instability better, it is essential to analyse more genomes with higher sequence coverage.

Repeat unit number variation can be a major contributor to pathogenic phenotypes, but it is still not fully understood. Many studies attempt to identify target genes and genetic instability characteristics that could be associated with carcinogenesis. In this analysis, I showed that when focusing only on the exonic regions of a genome mutational differences among tumor and healthy genomes need not differ in dramatic ways. Further investigation is needed to develop more precise and effective molecular diagnostic and therapeutic approaches.

7. References

Aaltonen, L.A., Peltomäki, P., Leach, F.S., Sistonen, P., Pylkkänen, L., Mecklin, J.P., Järvinen, H., Powell, S.M., Jen, J., Hamilton, S.R., 1993. Clues to the pathogenesis of familial colorectal cancer. Science 260, 812–816.

Beerenwinkel, N., Antal, T., Dingli, D., Traulsen, A., Kinzler, K.W., Velculescu, V.E., Vogelstein, B., Nowak, M.A., 2007. Genetic Progression and the Waiting Time to Cancer. PLoS Comput Biol 3, e225. doi:10.1371/journal.pcbi.0030225

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucl. Acids Res. 27, 573–580. doi:10.1093/nar/27.2.573

Boland, C.R., Goel, A., 2010. Microsatellite instability in colorectal cancer. Gastroenterology 138, 2073–2087.e3. doi:10.1053/j.gastro.2009.12.064

Boland, C.R., Thibodeau, S.N., Hamilton, S.R., Sidransky, D., Eshleman, J.R., Burt, R.W., Meltzer, S.J., Rodriguez-Bigas, M.A., Fodde, R., Ranzani, G.N., Srivastava, S., 1998. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res. 58, 5248–5257.

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., Rolf, B., 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62, 1408–1415.

Brook, J.D., McCurrach, M.E., Harley, H.G., Buckler, A.J., Church, D., Aburatani, H., Hunter, K., Stanton, V.P., Thirion, J.P., Hudson, T., 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell 68, 799–808.

Cancer Genome Atlas Network, 2012. Comprehensive molecular characterization of human colon and rectal cancer. Nature 487, 330–337. doi:10.1038/nature11252

Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J., Deka, R., 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. U.S.A. 94, 1041–1046.

Davies, H., Bignell, G.R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M.J., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., Menzies, A., Mould, C., Parker, A., Stevens, C., Watt, S., Hooper, S., Wilson, R., Jayatilake, H., Gusterson, B.A., Cooper, C., Shipley, J., Hargrave, D., Pritchard-Jones, K., Maitland, N., Chenevix-Trench, G., Riggins, G.J., Bigner, D.D., Palmieri, G., Cossu, A., Flanagan, A., Nicholson, A., Ho, J.W.C., Leung, S.Y., Yuen, S.T., Weber, B.L., Seigler, H.F., Darrow, T.L., Paterson, H., Marais, R., Marshall, C.J., Wooster, R., Stratton, M.R., Futreal, P.A., 2002. Mutations of the BRAF gene in human cancer. Nature 417, 949–954. doi:10.1038/nature00766

Denoeud, F., Vergnaud, G., Benson, G., 2003. Predicting Human Minisatellite Polymorphism. Genome Res 13, 856–867. doi:10.1101/gr.574403

Di Pietro, M., Sabates Bellver, J., Menigatti, M., Bannwart, F., Schnider, A., Russell, A., Truninger, K., Jiricny, J., Marra, G., 2005. Defective DNA mismatch repair determines a characteristic transcriptional profile in proximal colon cancers. Gastroenterology 129, 1047–1059. doi:10.1053/j.gastro.2005.06.028

Durbin, R.M., 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Gibbs, R.A., Hurles, M.E., McVean, G.A., 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073. doi:10.1038/nature09534

Duval, A., Hamelin, R., 2002a. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res. 62, 2447–2454.

Duval, A., Hamelin, R., 2002b. Genetic instability in human mismatch repair deficient cancers. Ann. Genet. 45, 71–75.

Ellegren, H., 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5, 435–445. doi:10.1038/nrg1348

Eyre, T.A., Ducluzeau, F., Sneddon, T.P., Povey, S., Bruford, E.A., Lush, M.J., 2006. The HUGO Gene Nomenclature Database, 2006 updates. Nucl. Acids Res. 34, D319–D321. doi:10.1093/nar/gkj147

Fan, H., Chu, J.-Y., 2007. A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics 5, 7–14. doi:10.1016/S1672-0229(07)60009-6

Fearon, E.R., 2011. Molecular genetics of colorectal cancer. Annu Rev Pathol 6, 479–507. doi:10.1146/annurev-pathol-011110-130235

Fearon, E.R., Vogelstein, B., 1990. A genetic model for colorectal tumorigenesis. Cell 61, 759–767.

Fondon, J.W., Garner, H.R., 2004. Molecular origins of rapid and continuous morphological evolution. Proc. Natl. Acad. Sci. U.S.A. 101, 18058–18063. doi:10.1073/pnas.0408118101

Fresno Vara, J.A., Casado, E., de Castro, J., Cejas, P., Belda-Iniesta, C., González-Barón, M., 2004. PI3K/Akt signalling pathway and cancer. Cancer Treat. Rev. 30, 193–204. doi:10.1016/j.ctrv.2003.07.007

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., Stratton, M.R., 2004. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183. doi:10.1038/nrc1299

Gemayel, R., Vinces, M.D., Legendre, M., Verstrepen, K.J., 2010. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu. Rev. Genet. 44, 445–477. doi:10.1146/annurev-genet-072610-155046

Giovannucci, E., Stampfer, M.J., Krithivas, K., Brown, M., Brufsky, A., Talcott, J., Hennekens, C.H., Kantoff, P.W., 1997. The CAG repeat within the androgen receptor gene and its relationship to prostate cancer. PNAS 94, 3320–3323.

Haiman, C.A., Hankinson, S.E., Spiegelman, D., De Vivo, I., Colditz, G.A., Willett, W.C., Speizer, F.E., Hunter, D.J., 2000. A tetranucleotide repeat polymorphism in CYP19 and breast cancer risk. Int. J. Cancer 87, 204–210. doi:10.1002/1097-0215(20000715)87:2<204::AID-IJC8>3.0.CO;2-3

Hanahan, D., Weinberg, R.A., 2000. The hallmarks of cancer. Cell 100, 57–70.

Harris, S.L., Levine, A.J., 2005. The p53 pathway: positive and negative feedback loops. Oncogene 24, 2899–2908. doi:10.1038/sj.onc.1208615

Imai, K., Yamamoto, H., 2008. Carcinogenesis and microsatellite instability: the interrelationship between genetics and epigenetics. Carcinogenesis 29, 673–680. doi:10.1093/carcin/bgm228

Ionov, Y., Peinado, M.A., Malkhosyan, S., Shibata, D., Perucho, M., 1993. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature 363, 558–561. doi:10.1038/363558a0

Jass, J.R., 2007. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 50, 113–130. doi:10.1111/j.1365-2559.2006.02549.x

Kan, Z., Jaiswal, B.S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H.M., Yue, P., Haverty, P.M., Bourgon, R., Zheng, J., Moorhead, M., Chaudhuri, S., Tomsho, L.P., Peters, B.A., Pujara, K., Cordes, S., Davis, D.P., Carlton, V.E.H., Yuan, W., Li, L., Wang, W., Eigenbrot, C., Kaminker, J.S., Eberhard, D.A., Waring, P., Schuster, S.C., Modrusan, Z., Zhang, Z., Stokoe, D., de Sauvage, F.J., Faham, M., Seshagiri, S., 2010. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466, 869–873. doi:10.1038/nature09208

Kashi, Y., King, D.G., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253–259. doi:10.1016/j.tig.2006.03.005

Kayser, M., de Knijff, P., 2011. Improving human forensics through advances in genetics, genomics and molecular biology. Nat. Rev. Genet. 12, 179–192. doi:10.1038/nrg2952

Kelkar, Y.D., Tyekucheva, S., Chiaromonte, F., Makova, K.D., 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18, 30–38. doi:10.1101/gr.7113408

Khan, R., Mittelman, D., 2013. Rumors of the death of consumer genomics are greatly exaggerated. Genome Biol. 14, 139. doi:10.1186/gb4141

Kit, S., 1961. Equilibrium Sedimentation in Density Gradients of Dna Preparations from Animal Tissues. J. Mol. Biol. 3, 711–&.

Knudson, A.G., 2002. Cancer genetics. Am. J. Med. Genet. 111, 96–102. doi:10.1002/ajmg.10320

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas,

P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. doi:10.1038/35057062

Legendre, M., Pochet, N., Pak, T., Verstrepen, K.J., 2007. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res. 17, 1787–1796. doi:10.1101/gr.6554007

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Lim, K.G., Kwoh, C.K., Hsu, L.Y., Wirawan, A., 2013. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. Brief. Bioinformatics 14, 67–81. doi:10.1093/bib/bbs023

Liquori, C.L., Ricker, K., Moseley, M.L., Jacobsen, J.F., Kress, W., Naylor, S.L., Day, J.W., Ranum, L.P., 2001. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science 293, 864–867. doi:10.1126/science.1062125

Logan, C.Y., Nusse, R., 2004. The Wnt signaling pathway in development and disease. Annu. Rev. Cell Dev. Biol. 20, 781–810. doi:10.1146/annurev.cellbio.20.010403.113126

López Castel, A., Cleary, J.D., Pearson, C.E., 2010. Repeat instability as the basis for human diseases and as a potential target for therapy. Nat Rev Mol Cell Biol 11, 165–170. doi:10.1038/nrm2854

Lovett, S.T., 2004. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. Mol. Microbiol. 52, 1243–1253. doi:10.1111/j.1365-2958.2004.04076.x

Madsen, B.E., Villesen, P., Wiuf, C., 2007. A periodic pattern of SNPs in the human genome. Genome Res. 17, 1414–1419. doi:10.1101/gr.6223207

Madsen, B.E., Villesen, P., Wiuf, C., 2008. Short tandem repeats in human exons: a target for disease mutations. BMC Genomics 9, 410. doi:10.1186/1471-2164-9-410

Mann, H.B., Whitney, D.R., 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann. Math. Statist. 18, 50–60. doi:10.1214/aoms/1177730491

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R.S., Zborowska, E., Kinzler, K.W., Vogelstein, B., Et, A., 1995. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science 268, 1336–1338. doi:10.1126/science.7761852

McIver, L.J., Fonville, N.C., Karunasena, E., Garner, H.R., 2014. Microsatellite genotyping reveals a signature in breast cancer exomes. Breast Cancer Res Treat 145, 791–798. doi:10.1007/s10549-014-2908-8

Näslund, K., Saetre, P., von Salomé, J., Bergström, T.F., Jareborg, N., Jazin, E., 2005. Genome-wide prediction of human VNTRs. Genomics 85, 24–35. doi:10.1016/j.ygeno.2004.10.009

Nowell, P.C., 2002. Tumor progression: a brief historical perspective. Semin. Cancer Biol. 12, 261–266.

O'Dushlaine, C.T., Edwards, R.J., Park, S.D., Shields, D.C., 2005. Tandem repeat copy-number variation in protein-coding regions of human genes. Genome Biology 6, R69. doi:10.1186/gb-2005-6-8-r69

O'Dushlaine, C.T., Shields, D.C., 2008. Marked variation in predicted and observed variability of tandem repeat loci across the human genome. BMC Genomics 9, 175. doi:10.1186/1471-2164-9-175

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

Payseur, B.A., Jing, P., Haasl, R.J., 2011. A genomic portrait of human microsatellite variation. Mol. Biol. Evol. 28, 303–312. doi:10.1093/molbev/msq198

Peltomäki, P., Aaltonen, L.A., Sistonen, P., Pylkkänen, L., Mecklin, J.P., Järvinen, H., Green, J.S., Jass, J.R., Weber, J.L., Leach, F.S., 1993. Genetic mapping of a locus predisposing to human colorectal cancer. Science 260, 810–812.

Perucho, M., 2003. Tumors with microsatellite instability: many mutations, targets and paradoxes. Oncogene 22, 2223–2225. doi:10.1038/sj.onc.1206580

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B., Gray, J.W., Albertson, D.G., 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20, 207–211. doi:10.1038/2524

Richard, G.F., Pâques, F., 2000. Mini- and microsatellite expansions: the recombination connection. EMBO Rep. 1, 122–126. doi:10.1038/sj.embor.embor606

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.-Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Etten, W.J.V., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S., Altshuler, D., 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409, 928–933. doi:10.1038/35057149

Sawyer, L.A., Hennessy, J.M., Peixoto, A.A., Rosato, E., Parkinson, H., Costa, R., Kyriacou, C.P., 1997. Natural variation in a Drosophila clock gene and temperature compensation. Science 278, 2117–2120.

Siegel, R., Ma, J., Zou, Z., Jemal, A., 2014. Cancer statistics, 2014. CA A Cancer Journal for Clinicians 64, 9–29. doi:10.3322/caac.21208

Siva, N., 2008. 1000 Genomes project. Nat. Biotechnol. 26, 256. doi:10.1038/nbt0308-256b

Stankiewicz, P., Lupski, J.R., 2010. Structural Variation in the Human Genome and its Role in Disease. Annual Review of Medicine 61, 437–455. doi:10.1146/annurev-med-100708-204735

Stern, A., Brown, M., Nickel, P., Meyer, T.F., 1986. Opacity genes in Neisseria gonorrhoeae: control of phase and antigenic variation. Cell 47, 61–71.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., Grassi, A. de, Lee, C., Tyler-Smith, C., Carter, N., Scherer, S.W., Tavaré, S., Deloukas, P., Hurles, M.E., Dermitzakis, E.T., 2007. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science 315, 848–853. doi:10.1126/science.1136678

The Huntington's Disease Collaborative Research Group, 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 72, 971–983.

Thibodeau, S.N., Bren, G., Schaid, D., 1993. Microsatellite instability in cancer of the proximal colon. Science 260, 816–819.

Tóth, G., Gáspári, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.

UK, C.R., 2014. Worldwide cancer statistics [WWW Document]. URL http://www.cancerresearchuk.org/cancer-info/cancerstats/world/ (accessed 11.17.14).

Verkerk, A.J., Pieretti, M., Sutcliffe, J.S., Fu, Y.H., Kuhl, D.P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F.P., 1991. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65, 905–914.

Verstrepen, K.J., Jansen, A., Lewitter, F., Fink, G.R., 2005. Intragenic tandem repeats generate functional variability. Nat. Genet. 37, 986–990. doi:10.1038/ng1618

Vilar, E., Gruber, S.B., 2010. Microsatellite instability in colorectal cancer-the stable evidence. Nat Rev Clin Oncol 7, 153–162. doi:10.1038/nrclinonc.2009.237

Vogelstein, B., Fearon, E., Hamilton, S., Kern, S., Preisinger, A., Leppert, M., Nakamura, Y., White, R., Smits, A., Bos, J., 1988. Genetic Alterations During Colorectal-Tumor Development. N. Engl. J. Med. 319, 525–532. doi:10.1056/NEJM198809013190901

Vogelstein, B., Kinzler, K.W., 2004. Cancer genes and the pathways they control. Nat Med 10, 789–799. doi:10.1038/nm1087

Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., Lander, E.S., 1998. Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. Science 280, 1077–1082. doi:10.1126/science.280.5366.1077

Weber, J.L., Wong, C., 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. 2, 1123–1128.

Wilks, C., Cline, M.S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J., Nelson, D., Litzinger, B., Hatton, T., Maltbie, L., Ainsworth, M., Allen, P., Rosewood, L., Mitchell, E., Smith, B., Warner, J., Groboske, J., Telc, H., Wilson, D., Sanford, B., Schmidt, H., Haussler, D., Maltbie, D., 2014. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database 2014, bau093–bau093. doi:10.1093/database/bau093

Willems, T.F., Gymrek, M., Highnam, G., Mittelman, D., Erlich, Y., 2014. The landscape of human STR variation. Genome Res. gr.177774.114. doi:10.1101/gr.177774.114

Woerner, S.M., Benner, A., Sutter, C., Schiller, M., Yuan, Y.P., Keller, G., Bork, P., Doeberitz, M. von K., Gebert, J.F., 2003. Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. Oncogene 22, 2226–2235. doi:10.1038/sj.onc.1206421

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P.A., Kaminker, J.S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J.K.V., Sukumar, S., Polyak, K., Park, B.H., Pethiyagoda, C.L., Pant, P.V.K., Ballinger, D.G., Sparks, A.B., Hartigan, J., Smith, D.R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S.D., Parmigiani, G., Kinzler, K.W., Velculescu, V.E., Vogelstein, B., 2007. The genomic landscapes of human breast and colorectal cancers. Science 318, 1108–1113. doi:10.1126/science.1145720

Zhao, J., Grant, S.F.A., 2011. Advances in whole genome sequencing technology. Curr Pharm Biotechnol 12, 293–305.

8. Supplementary Material

patient	cancer type	patient	cancer type	patient	cancer type
A6-2671	Colon	AA-3968	Colon	AF-2691	Rectum
A6-2676	Colon	AA-A00U	Colon	AF-2692	Rectum
A6-2678	Colon	AA-A00Z	Colon	AG-3574	Rectum
A6-2680	Colon	AA-A01K	Colon	AG-3728	Rectum
A6-2681	Colon	AA-A01P	Colon	AG-3878	Rectum
A6-3807	Colon	AA-A01R	Colon	AG-3887	Rectum
AA-3495	Colon	AA-A01T	Colon	AG-3892	Rectum
AA-3502	Colon	AA-A01V	Colon	AG-3894	Rectum
AA-3514	Colon	AA-A02K	Colon	AG-3902	Rectum
AA-3516	Colon	AA-A02R	Colon	AG-3909	Rectum
AA-3529	Colon	AZ-4315	Colon	AG-4001	Rectum
AA-3548	Colon	AZ-4681	Colon	AG-4005	Rectum
AA-3549	Colon	AZ-4682	Colon	AG-4007	Rectum
AA-3555	Colon	AZ-4684	Colon	AG-4008	Rectum
AA-3558	Colon	CA-5256	Colon	AG-4015	Rectum
AA-3664	Colon	CK-4951	Colon	AG-A002	Rectum
AA-3666	Colon	CM-4746	Colon	AG-A00Y	Rectum
AA-3675	Colon	CM-4747	Colon	AG-A011	Rectum
AA-3685	Colon	CM-4748	Colon	AG-A01W	Rectum
AA-3861	Colon	CM-4750	Colon	AG-A01Y	Rectum
AA-3947	Colon	CM-4752	Colon	AG-A020	Rectum
AA-3956	Colon			AG-A032	Rectum

Supplementary Table S1. List of genomes considered in the study (n=62). The left column indicates TCGA (Cancer Genome Atlas Network, 2012) sequence IDs, the right column cancer type (colon or rectal)

pathway	gene	pathway	gene	pathway	gene
Wnt	APC2	mTOR	PDK2	p53	BCL2L14
Wnt	APCDD1L	mTOR	PDK3	p53	BCL2L15
Wnt	APCDD1	mTOR	PDK4	p53	BCL2L1
Wnt	APCS	mTOR	AKT1	p53	BCL2L2
Wnt	APC	mTOR	AKT2	p53	BCL2
Wnt	WNT10A	mTOR	AKT3	p53	CCNE1
Wnt	WNT10B	mTOR	STK11	p53	CCNE2
Wnt	WNT11	TGF_beta	TGFA	p53	CCND1
Wnt	WNT16	TGF_beta	TGFB1I1	p53	CCND2
Wnt	WNT1	TGF_beta	TGFB1	p53	CCND3

Wnt	WNT2B	TGF_beta	TGFB2	MAPK	JUN
Wnt	WNT2	TGF_beta	TGFB3	MAPK	PDCD4
Wnt	WNT3A	TGF_beta	TGFBI	MAPK	MAPK10
Wnt	WNT3	TGF_beta	TGFBR1	MAPK	MAPK11
Wnt	WNT4	TGF_beta	TGFBR2	MAPK	MAPK12
Wnt	WNT5A	TGF_beta	TGFBR3	MAPK	MAPK13
Wnt	WNT5B	TGF_beta	TGFBRAP1	MAPK	MAPK14
Wnt	WNT6	TGF_beta	SMAD1	MAPK	MAPK15
Wnt	WNT7A	TGF_beta	SMAD2	MAPK	MAPK1
Wnt	WNT7B	TGF_beta	SMAD3	MAPK	MAPK3
Wnt	WNT8A	TGF_beta	SMAD4	MAPK	MAPK4
Wnt	WNT8B	TGF_beta	SMAD5	MAPK	MAPK6
Wnt	WNT9A	TGF_beta	SMAD6	MAPK	MAPK7
Wnt	WNT9B	TGF_beta	SMAD7	MAPK	MAPK8
Wnt	CTNNB1	TGF_beta	SMAD9	MAPK	MAPK9
Wnt	CTNNBL1	TGF_beta	TNFRSF1A	MAPK	MAPKAP1
Wnt	AXIN1	TGF_beta	TNFRSF1B	MAPK	MAPKBP1
Wnt	AXIN2	TGF_beta	ACVR1B	MAPK	MAP2K1
Wnt	GSK3A	TGF_beta	ACVR1C	MAPK	MAP2K2
Wnt	GSK3B	TGF_beta	ACVR1	MAPK	MAP2K3
Wnt	BTRC	TGF_beta	ACVR2A	MAPK	MAP2K4
Wnt	CSNK1A1L	TGF_beta	ACVR2B	MAPK	MAP2K5
Wnt	CSNK1A1	TGF_beta	ACVRL1	MAPK	MAP2K6
Wnt	CSNK1D	TGF_beta	BMPR1A	MAPK	MAP2K7
Wnt	CSNK1E	TGF_beta	BMPR1B	MAPK	MAP3K10
Wnt	CSNK1G1	TGF_beta	HRAS	MAPK	MAP3K11
Wnt	CSNK1G2	p53	TP53AIP1	MAPK	MAP3K12
Wnt	CSNK1G3	p53	TP53BP1	MAPK	MAP3K13
Wnt	DVL1	p53	TP53BP2	MAPK	MAP3K14
Wnt	DVL2	p53	TP53I11	MAPK	MAP3K15
Wnt	DVL3	p53	TP53I13	MAPK	MAP3K1
Wnt	TCF12	p53	TP53I3	MAPK	MAP3K2
Wnt	TCF15	p53	TP53INP1	MAPK	MAP3K3
Wnt	TCF21	p53	TP53INP2	MAPK	MAP3K4
Wnt	TCF23	p53	TP53RK	MAPK	MAP3K5
Wnt	TCF25	p53	TP53TG1	MAPK	MAP3K6
Wnt	TCF3	p53	TP53TG3B	MAPK	MAP3K7
Wnt	TCF4	p53	TP53TG5	MAPK	MAP3K8
Wnt	TCF7L1	p53	TP53	MAPK	MAP3K9
Wnt	TCF7L2	p53	MDM2	MAPK	MAP4K1
Wnt	TCF7	p53	ATMIN	MAPK	MAP4K2
Wnt	TCFL5	p53	ATM	MAPK	MAP4K3

Wnt	TLE1	p53	CASP10	MAPK	MAP4K4
Wnt	TLE2	p53	CASP12	MAPK	MAP4K5
Wnt	TLE3	p53	CASP14	MAPK	MAP4
Wnt	TLE4	p53	CASP1	MAPK	MAP6D1
Wnt	TLE6	p53	CASP2	MAPK	MAP6
Wnt	CREBBP	p53	CASP3	MAPK	MAP7D1
Wnt	EP300	p53	CASP4	MAPK	MAP7D2
Wnt	LRP10	p53	CASP5	MAPK	MAP7D3
Wnt	LRP11	p53	CASP6	MAPK	MAP7
Wnt	LRP12	p53	CASP7	MAPK	MAP9
Wnt	LRP1B	p53	CASP8	MAPK	KRAS
Wnt	LRP1	p53	CASP9	MAPK	BRAF
Wnt	LRP4	p53	FASLG	MAPK	NRAS
Wnt	LRP5L	p53	FASN	MAPK	EGFR
Wnt	LRP5	p53	FASTKD1	MAPK	ERBB2
Wnt	LRP6	p53	FASTKD2	MAPK	ERBB3
Wnt	LEF1	p53	FASTKD3	MAPK	ERBB4
Wnt	MT1B	p53	FASTKD5	MAPK	FGF10
Wnt	NKD1	p53	FASTK	MAPK	FGF11
Wnt	NKD2	p53	FAS	MAPK	FGF12
Wnt	DKK1	p53	CDC20B	MAPK	FGF13
Wnt	DKK2	p53	CDC20	MAPK	FGF14
Wnt	DKK3	p53	CDC23	MAPK	FGF16
Wnt	DKK4	p53	CDC25A	MAPK	FGF17
Wnt	CTBP1	p53	CDC25B	MAPK	FGF18
Wnt	CTBP2	p53	CDC25C	MAPK	FGF19
Wnt	SFRP1	p53	CDC26	MAPK	FGF1
Wnt	SFRP2	p53	CDC27	MAPK	FGF20
Wnt	SFRP4	p53	BAX	MAPK	FGF21
Wnt	SFRP5	p53	NOXA1	MAPK	FGF22
Wnt	RHOA	p53	BBC3	MAPK	FGF23
Wnt	RTKN2	p53	CHEK1	MAPK	FGF2
Wnt	RTKN	p53	CHEK2	MAPK	FGF3
Wnt	CDX2	p53	SIRT1	MAPK	FGF4
Wnt	FBXW2	p53	CDK10	MAPK	FGF5
mTOR	PIP4K2A	p53	CDK11A	MAPK	FGF6
mTOR	PIP4K2B	p53	CDK11B	MAPK	FGF7
mTOR	PIP4K2C	p53	CDK12	MAPK	FGF8
mTOR	PIP5K1A	p53	CDK13	MAPK	FGF9
mTOR	PIP5K1B	p53	CDK14	MAPK	FGFR1
mTOR	PIP5K1C	p53	CDK15	MAPK	FGFR2
mTOR	PIP5K1P1	p53	CDK16	MAPK	FGFR3

mTOR	PIP5KL1	p53	CDK17	MAPK	FGFRL1
mTOR	PIPOX	p53	CDK18	MAPK	MYC
mTOR	PIPSL	p53	CDK19	MAPK	RAF1
mTOR	PIP	p53	CDK1	MAPK	RASA1
mTOR	PTENP1	p53	CDK20	MAPK	RASA2
mTOR	PTEN	p53	CDK2	MAPK	RASA3
mTOR	MTOR	p53	CDK3	MAPK	RASA4
mTOR	IGF1R	p53	CDK4	MAPK	RASD1
mTOR	IGF1	p53	CDK5	MAPK	RASD2
mTOR	IGF2R	p53	CDK6	MAPK	RASEF
mTOR	IGF2	p53	CDK8	MAPK	RASGEF1A
mTOR	IRS1	p53	CDK9	MAPK	RASGEF1B
mTOR	IRS2	p53	CDKL1	MAPK	RASGEF1C
mTOR	IRS4	p53	CDKL2	MAPK	RASGRF1
mTOR	PIK3AP1	p53	CDKL3	MAPK	RASGRF2
mTOR	PIK3C2A	p53	CDKL4	MAPK	RASGRP1
mTOR	PIK3C2B	p53	CDKL5	MAPK	RASGRP2
mTOR	PIK3C2G	p53	CDKN1A	MAPK	RASGRP3
mTOR	PIK3C3	p53	CDKN1B	MAPK	RASGRP4
mTOR	PIK3CA	p53	CDKN1C	MAPK	PRKAA1
mTOR	PIK3CB	p53	CDKN2A	MAPK	PRKAA2
mTOR	PIK3CD	p53	CDKN2B	MAPK	PRKAB1
mTOR	PIK3CG	p53	CDKN2C	MAPK	PRKAB2
mTOR	PIK3R1	p53	CDKN2D	MAPK	PRKACA
mTOR	PIK3R2	p53	CDKN3	MAPK	PRKACB
mTOR	PIK3R3	p53	BCL2A1	MAPK	PRKACG
mTOR	PIK3R4	p53	BCL2L10	MAPK	PRKAG1
mTOR	PIK3R5	p53	BCL2L11	MAPK	PRKAG2
mTOR	PIK3R6	p53	BCL2L12	MAPK	PRKAG3
mTOR	PDK1	p53	BCL2L13		

Supplementary Table S2. List of genes involved in cancer pathways (n=371), indicated with their HUGO (Eyre et al. 2006) gene IDs and the patways (Bilgin Sonay et al., 2015) they are associated with.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Comprehensive characterization of tandem repeat instability in 62 colorectal tumors.

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Koletou	Malamati

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
ZURICH, 15/12/2014	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.