Human somatic mutation rate estimated from SNP chips

Running title

Human somatic mutation rate

Author

Matthias Wjst

Molekulare Epidemiologie

GSF – Forschungszentrum für Umwelt und Gesundheit

Ingolstädter Landstrasse 1

D-85764 Neuherberg, Germany

Tel.: ++49-89-3187-4565

Fax: ++49-89-3187-3533

E-Mail: wjst@gsf.de

Revision 22-June-2007

Abstract

There is substantial phylogenetic and epidemiological information about the distribution of SNPs in the human population while there is only limited information how SNPs are introduced into an individual genome.

Here I show in a genome-wide panel of SNPs that children have multiple variant sites that are not present in their parents. The median mutation rate is unexpected high with 10,6 mutations / Mb /diploid genome in Yorubean (11,9 in European) children even after applying thorough rules for genotype quality. Ex vivo artefacts may have contributed to false positive counts, however, mutation sites do not match known cell culture aberrations. Scanning preferentially mutation-prone SNP sites in DNA repeats or retroviral inserts by current genomewide SNP panles seemed to be the main reason for the high mutation rate.

Further studies using whole genome sequencing may yield more exact mutation rates - a finding that could have implications in forensic science as well as in ageing and cancer research.

Background

Somatic mutations are thought to accumulate during human lifetime (Boland and Goel, 2005), (Lou and Chen, 2006), (Crow, 2006) either as errors during DNA replication that escape DNA repair (Rubin, 2006) or by direct oxidative or radiation induced DNA damage (Shigenaga, et al., 1994). While ageing has been previously defined as a telomere function (Wong, et al., 2003) accumulation of DNA damage leading to stem cell exhaustion is now a major research focus (Nijnik, et al., 2007).

Normal human somatic mutation rate has been estimated to be in the range between 0,02 and 25 mutations / Mb / generation (see online supplement sheet A and references (Wong, et al., 2003), (Oller, et al., 1989), (Weinberg, et al., 2001), (Kondrashov, 2003) and (Greenman, et al., 2007))) with the majority of studies describing less than 1 mutation / Mb / generation. Numbers, however, are difficult to compare as being obtained by diverse methods, in various source tissues and using different extrapolations, making any firm conclusions difficult.

Mutations may be acquired during early germline development in parents or after fertilization during successive somatic division (Fig.1). In humans there are approximately 500 germline cell divisions leading to about 100 additional mutations transmitted to the next generation (Crow, 2006). The number of somatic division in a human is not known but expected to be approximately 5,000 (Potten, et al., 2002); a 80-year-old adult may have therefore acquired up to 15,000 mutations (Crow, 1997). In constrast, 6,000 to 11,000 genomic alterations have been estimated to occur in a single cancer cell (Greenman, et al., 2007), (Stoler, et al., 1999), (Wang, et al., 2002) which comes close to the general estimates above.

The technical capability to determine genome-wide mutation rate was limited in the past. Human mutation rates could be of forensic interest when dealing with unknown DNA samples. Other potential areas of interest are the study of the normal ageing process in different tissues (Weiss, 2005) and early events

leading to cancer (Suh and Vijg, 2006), (Sjoblom, et al., 2006).

To gain a further insight into human mutation rates, I have chosen a public domain dataset of genome-wide SNP genotypes that was obtained by the Affymetrix GeneChip Mapping 500K Array Set. It uses two arrays, each capable of genotyping on average 250,000 SNPs (262,000 for Nsp1 arrays and 238,000 for Sty1 arrays). Nsp1 and Sty1 denote the initial restriction enzyme cutting of the native DNA, before it is being ligated, amplified, fragmented, labelled and hybridized to a chip that is finally scanned for fluorescence signals. The fluorescence signals were further processed (NN, 2006) and finally assigned a genotype (AA, AB or BB) including a quality score that integrates signal characteristics.

Methods

Individuals contributing DNA samples to the Hapmap project (NN, 2005) have come from a total of 270 people and 4 ethnically diverse groups. The Yoruba people of Ibadan, Nigeria, provided 30 sets of samples from two parents and one child (YRI) similar to 30 further trios collected in 1980 from U.S. residents with Northern and Western European ancestry by the Centre d'Etude du Polymorphisme Humain registry (CEU). The remaining samples of 45 Chinese (CHB) and 45 Japanese (JPT) were not include here as the donors were not related.

Following sample interrogation genotypes were called by the Affymetrix GeneChip Genotyping Analysis Software (GTYPE). Data were downloaded from the public Affymetrix website (www.affymetrix.com) in Jan 2006 and again in December 2006 including 500,568 SNP genotypes for each individual. The initial analysis was done with the DM calling algorithm but was replaced later by the BRLMM method (Rabbee and Speed, 2006). All analysis has been done using R statistical software 2.1.0.

Results

The table (online supplement sheet B) lists mutation counts in all children as "AB" calls where either both parents have "AA" genotypes or both parents have "BB" genotypes. Mutation counts are exceptional high (Fig. 2).

Unfortunately true mutations can not be unequivocally identified in this dataset due to the unknown fraction of genotyping errors (NN, 2006). As non-paternity may be a particular "error trap", I looked at the corresponding quality scores. The scores of the de novo mutations have a largely different distribution compared to "regular" calls at the same sites (Fig.3). When restricting therefore genotype calls to those with scores of <0.1 (the turning point between both distributions), the number of mutations drops to probably more realistic estimates (online supplement sheet B) – in particular when the restriction to "good" scores is being applied also to both parental genotypes.

A verification by resequencing (Green, et al., 2006) seems to be necessary but does not make too much sense as will be discussed later. At present, it is not fully clear if genotypes with poorer scores are indeed false positive counts as they may include also somatic mosaics (where the mutation is still not present in all cells) that may lead to difficulties in the scoring algorithm and hence less reliable scores.

By using a score threshold of <0,1 the corrected mutation rate is estimated to be 10,6 mutations / Mb diploid genome / lifetime in Yorubean (and 11,9 in the European population (online supplement sheet B). This estimate isl higher than in cancer tissues (Greenman, et al., 2007) and there are even outliers with high mutation oounts (online supplement sheet B). I therefore tried to characterize the newly found mutations in more detail. It seems that they are evenly distributed over the human genome (Fig.4) with only 4 cluster observed when scanning moving windows of 50 kB distance.

Cluster #1 contained 6 diverse SNPs (rs764243, rs3120697, rs3129567, rs3120699, rs6426311, rs3124124 and rs10802426) that are all intronic to

TFB2M. Cluster #2 was found at rs1340615, where 7 mutations are observed in a gene desert. Cluster 3 shows 6 SNPs (rs2240826, rs1546834, rs1546833, rs740821, rs2240834, rs2240835 and rs1860519) in TCRγ alternate reading frame protein. Cluster encompassed 2 frequently mutated SNPs (rs17004107 and rs7279082) both in a low conserved desert.

Nearly all cluster are situated in genomic regions of retroviral inserts or in DNA repeats (van den Hurk, et al., 2007). The mutations observed here do not match any known mutation hotspots (Rogozin, et al., 2001) nor are they being over-represented in cancer genes (Sjoblom, et al., 2006), (Futreal, et al., 2004), (Bamford, et al., 2004). Also a functional relevance of these mutations remains speculative as only a few SNPs are located in functionally relevant site (online supplement sheet C)

In a last step, I tried to follow up the individual history of the mutations if being introduced already during germline or during somatic development. The number of X chromosomal mutations in girls (XX) was higher than the double number of X chromosomal mutations in boys (XY)which is not unexpected as their single X chromosome did not undergo the extended lifecycle in the paternal germline. As germline cells are temporarily methylated (rendering them hypermutable), the number of autosomal CpG mutations correlated with the X chromosomal mutations in girls while no such correlation existed in boys. Unfortunately, the absolute number of included autosomal CpG sites with a score <0.1 was too low to allow for a formal testing.

Discussion

In a genome-wide panel of SNPs an excess of newly introduced somatic mutations was found in children that were otherwise not present in their parents.

Although it is generally assumed that BRLMM (the standard Affymetrix

algorithm) has an error rate under 0.2% (Consortium, 2007) a recent study of ~17,000 samples using the same SNP arrays excluded additionally 6.2% of all genotypes following extensive quality checks (Consortium, 2007). The increased threshold used here for quality scores leads to the exclusion of only ~86 SNPs per array which is a rather negligible number.

Reasons for the high number of mutations observed here are largely unknown. Although the possibility exists that recent studies have underestimated human mutation rates (the largest study so far (Greenman, et al., 2007) described also in non cancer human cell lines 5.6 mutations / Mb), the results here seemed to be biased for scanning mutation prone sites (as indicated by variants found in at least one human population before). The relationship of the mutation cluster to DNA repeats and retroposition elements may point towards a most likely explanation.

When did these mutations occur? Early germline mutations may be found in all body cells (Ellegren, 2002) while more recent mutations may not have reached the "100%" fixation rate. This may be concluded from studies of microchimerism where cells can be traced even decades later (Maloney, et al., 1999). By using 250 ng of genomic DNA as template for the polymerase chain reaction (as in the current setup) ~35000 single DNA molecules are used as a starting point. From previous studies it is possible that differences of <3% allele frequency may be discriminated by chip-based hybridization arrays (Kirov, et al., 2006), (Meaburn, et al., 2006), (Pearson, et al., 2007).

The kinetics of a single mutation, however, are impossible to examine as there is no possibility to label early human primordial cells and follow their progeny during the lifetime (although promising first experiments are available in animal models (Clayton, et al., 2007)). Monitoring involuntary radiation exposure (Weinberg, et al., 2001) or clonal analysis of repopulating cells after bone marrow transplantation (McKenzie, et al., 2006) has also shortcomings as these are either selected cell types or background conditions make mutation counts unreliable. There could be even a high individual

variation in the rate of mutational events as mutation of master switches like *ras* – may itself lead to hypo- (or hyper) methylation with consecutive over expression of further oncogenes (Boland and Goel, 2005) – a potential mechanism for some outliers in this dataset with a high rate of somatic mutations.

further resequencing of the current Hapmap samples is not warranted as the cell culture passage of immortalized lymphocytes has already introduced artefacts (Redon, et al., 2006). During karyotyping of Hapmap cell lines, several chromosomal abnormalities have been detected which makes it likely that there could be even more *ex vivo* mutations (Redon, et al., 2006).

Conclusion

It seems promising to test larger (probably 3 generational) families for somatic mutations (Youssoufian and Pyeritz, 2002). A main problem to solve is the distinction of genotyping errors and somatic mosaics to allow an unconditional recognition of mutations. Further work will most likely require repeated sequencing of the same individuals as only this approach will clearly delineate the time course of mutational events. With these first results on the birth and decay of SNPs at hand, it is tempting to speculate that refined mutation rates could lead to valid estimates of the "genomic age" of an individual (Nijnik, et al., 2007).

Authors Contribution

I developed the concept, did the analysis and wrote the paper.

Acknowledgments

I wish to acknowledge comments by Michael Knapp, Peter Reitmeir, Theresa Faus-Kessler and Michael Wittig on earlier versions of the manuscript as well as Dirk Jürgensen for help with analysis of the Affymetrix arrays.

Funding

I am funded by GSF FE 73922.

Conflicts of interest

I declare that I do not have conflicts of interest with the publication of this paper.

Figure 1: DNA transmission a across 3 generations. Starting with primordial (P) and stem (S) cells either differentiated (D), sperm (SP) and zygote (Z) cells develop. D cells may be seen as the result of symmetric self-renewing or expansion divisions (early P->S), asymmetric or maintenance divisions (early S->D) or differentiating divisions (terminal S->D). The number of cells in the various phases is tissue dependent and large unpredictable at a single cell level as the development follows a stochastic process. Actively dividing cells may even become quiescent, reactivated only at a later time, or die prematurely. Mutations may occur at any time – and dependent on the lifecycle of their origin – been found at DI, DII or DIII.

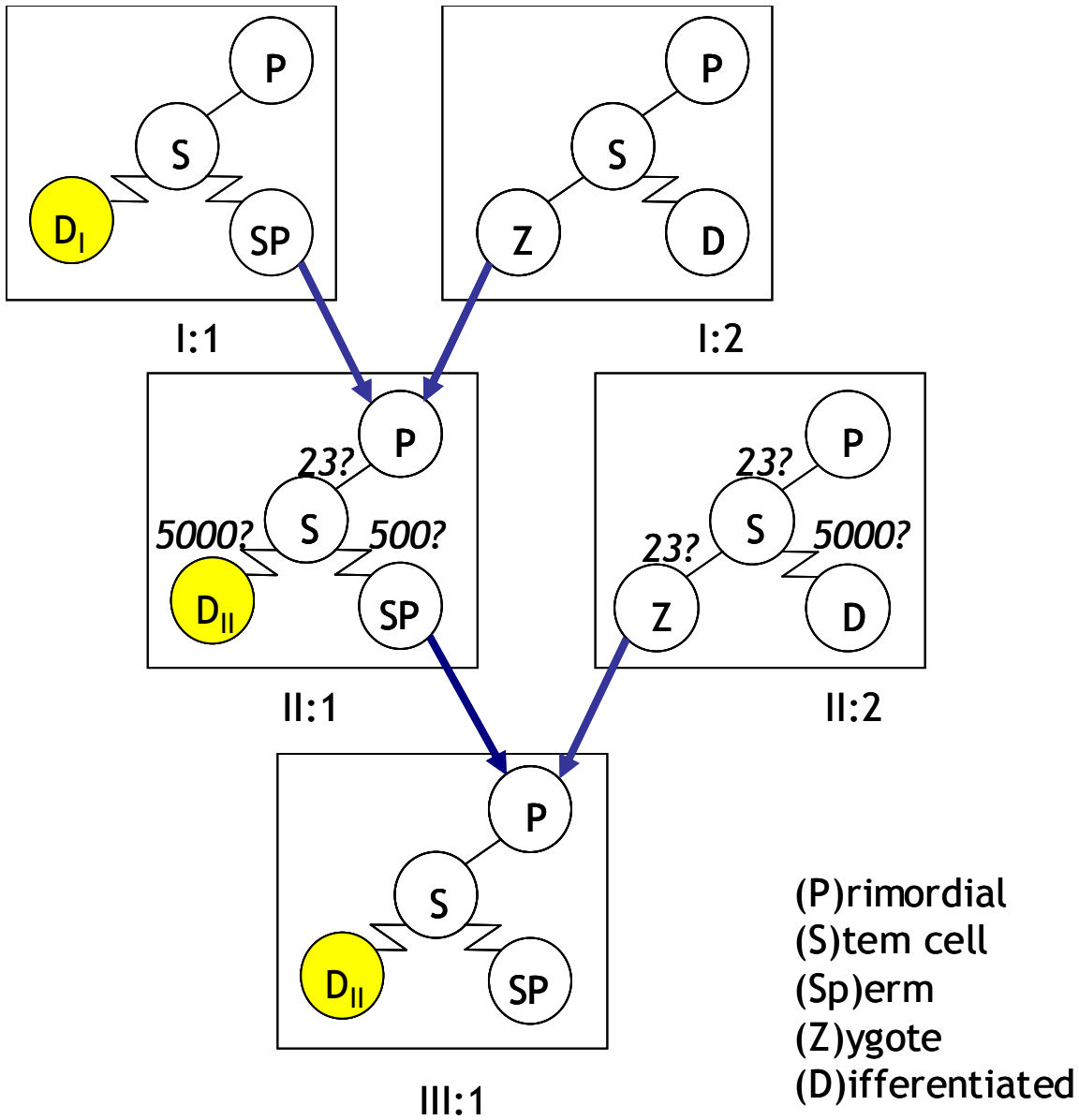


Figure 2: Crude count of acquired somatic mutations in 30 CEU children (white bars) and 30 YRI children (dark bars). Children were ordered from left to right by increasing number of observed mutations.

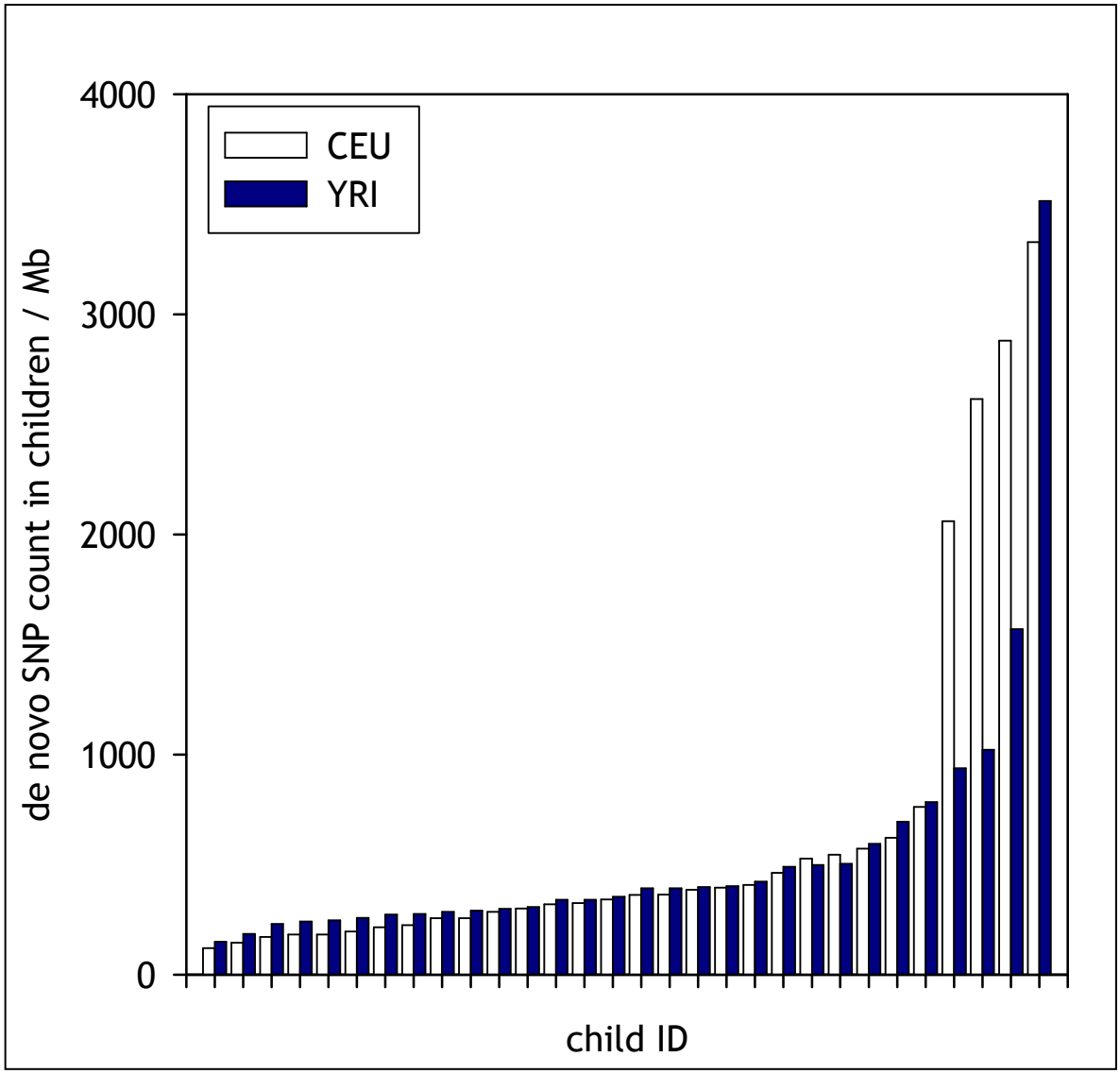


Figure 3: Distribution of genotype quality scores of two SNP assays (NAI10838 /left as a random example and NAI07348 /right as a SNP with a high count of somatic mutations). Somatic mutations have a different distribution of quality scores than regular calls with change point at 0.1.

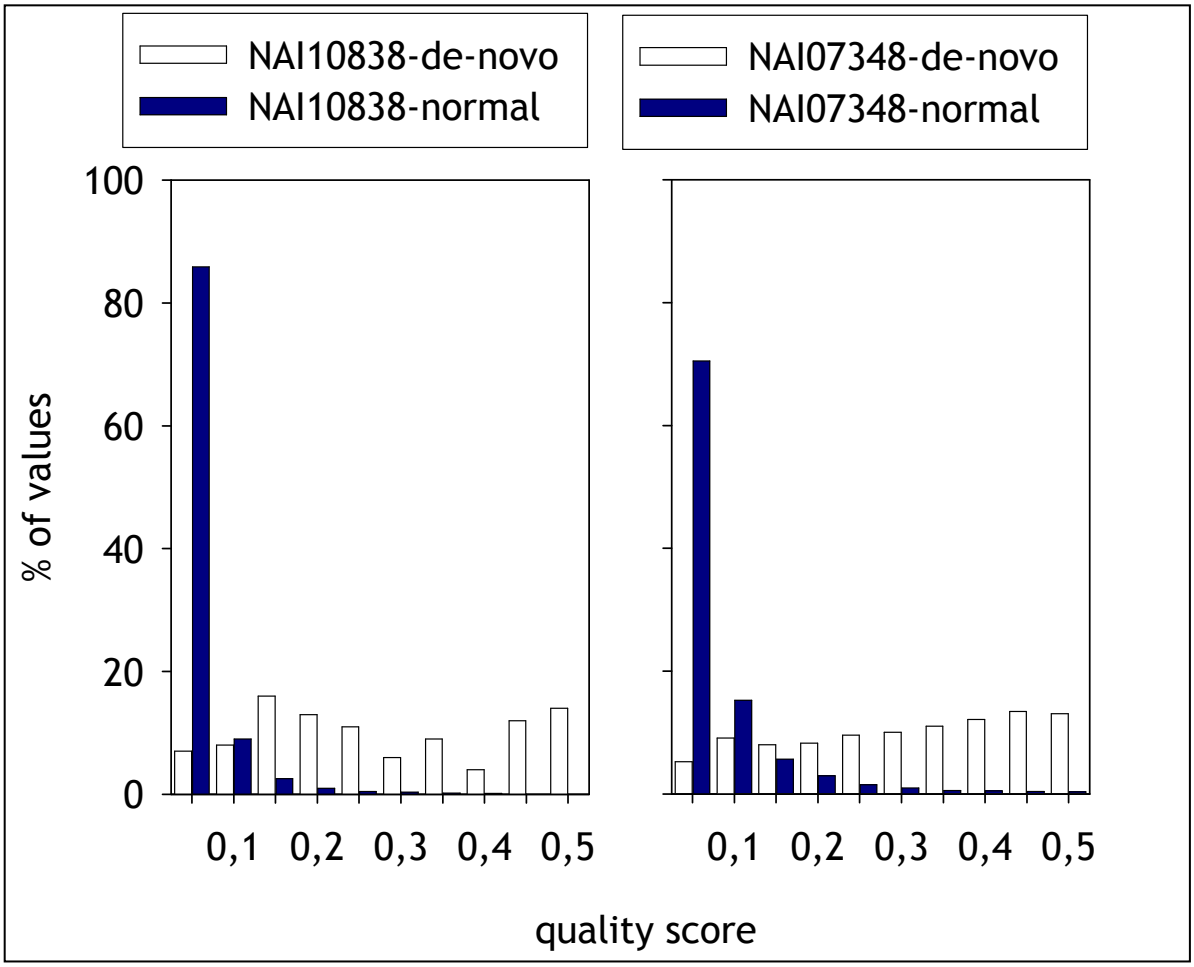
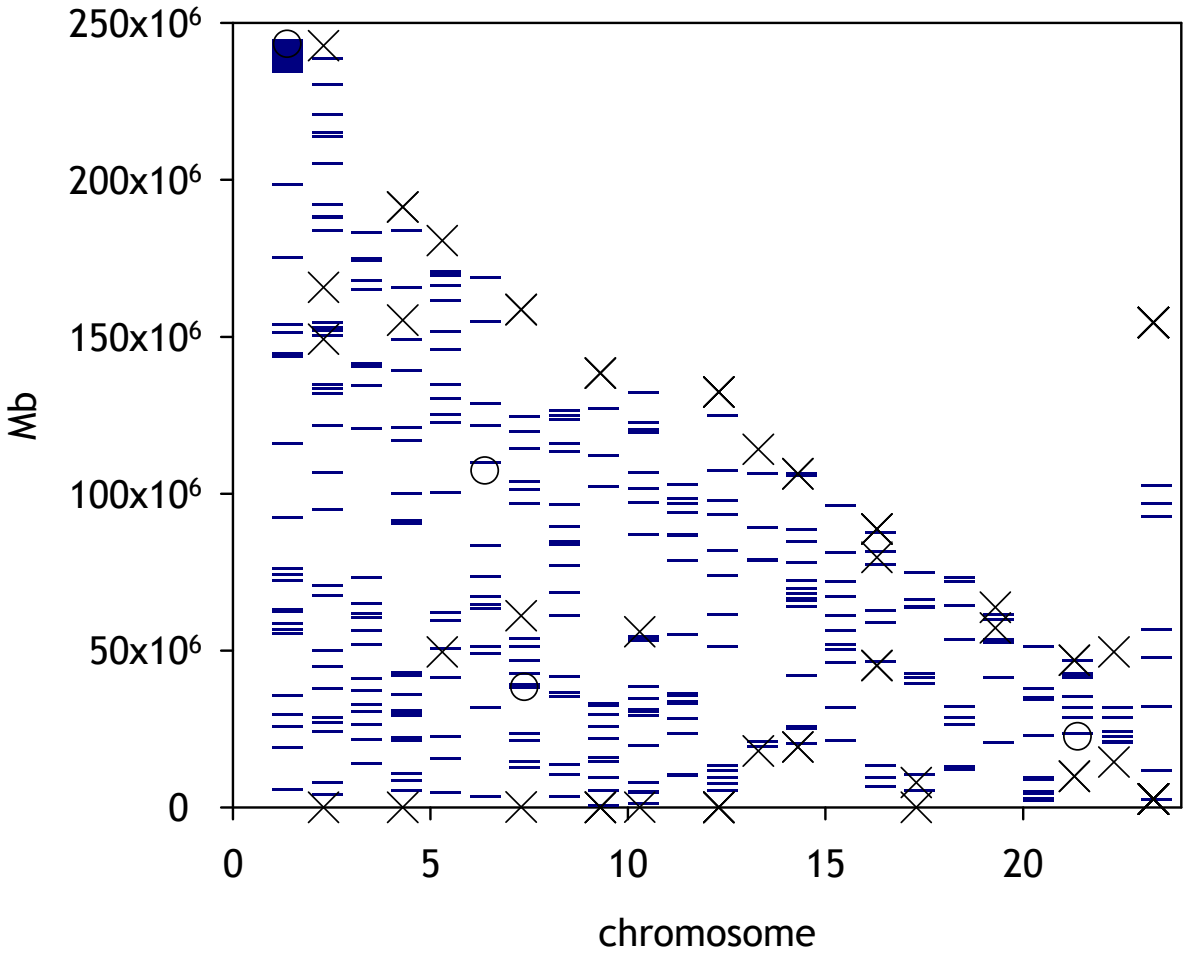


Figure 4: Location of all somatic mutations depicted as horizontal line with chromosome numbers increasing from left to right (X chromosome is shown as chromosome 23). Previously identified areas with chromosomal aberrations in cell lines (Redon, et al., 2006) are indicated by crosses. The four mutation clusters identified here are indicated by circles.



References

Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, Flanagan A, Teague J, Futreal PA, Stratton MR and others. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 91(2):355-8.

Boland CR, Goel A. 2005. Somatic evolution of cancer cells. Semin Cancer Biol 15(6):436-50.

Clayton E, Doupe DP, Klein AM, Winton DJ, Simons BD, Jones PH. 2007. A single type of progenitor cell maintains normal epidermis. Nature 446(7132):185-9.

Consortium WTCC. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145):661-78.

Crow JF. 1997. The high spontaneous mutation rate: Is it a health risk? Proc Natl Acad Sci U S A 94:8380-8386.

Crow JF. 2006. Age and sex effects on human mutation rates: an old problem with new complexities. J Radiat Res (Tokyo) 47 Suppl B:B75-82.

Ellegren H. 2002. Human mutation--blame (mostly) men. Nat Genet 31(1):9-10.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. Nat Rev Cancer 4(3):177-83.

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M and others. 2006. Analysis of one million base pairs of Neanderthal DNA. Nature 444(7117):330-6.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C and others. 2007. Patterns of somatic mutation in human cancer genomes. Nature 446(7132):153-8.

Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'Donovan M C. 2006. Pooled DNA genotyping on Affymetrix SNP genotyping arrays. BMC Genomics 7(1):27.

Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21(1):12-27.

Lou Z, Chen J. 2006. Cellular senescence and DNA repair. Exp Cell Res 312(14):2641-6.

Maloney S, Smith A, Furst DE, Myerson D, Rupert K, Evans PC, Nelson JL. 1999. Microchimerism of maternal origin persists into adult life. J Clin Invest

104(1):41-7.

McKenzie JL, Gan OI, Doedens M, Wang JC, Dick JE. 2006. Individual stem cells with highly variable proliferation and self-renewal properties comprise the human hematopoietic stem cell compartment. Nat Immunol 7(11):1225-33.

Meaburn E, Butcher LM, Schalkwyk LC, Plomin R. 2006. Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. Nucleic Acids Res 34(4):e27.

Nijnik A, Woodbine L, Marchetti C, Dawson S, Lambe T, Liu C, Rodrigues NP, Crockford TL, Cabuy E, Vindigni A and others. 2007. DNA repair is limiting for haematopoietic stem cells during ageing. Nature 447(7145):686-90.

NN. 2005. A haplotype map of the human genome. Nature 437(7063):1299-320.

NN. 2006. BRLMM: an improved genotype calling method for the GeneChip(R) Human Mapping 500K Array Set. Affymetrix Whitepaper 1.0 / 2006-04-14.

Oller AR, Rastogi P, Morgenthaler S, Thilly WG. 1989. A statistical model to estimate variance in long term-low dose mutation assays: testing of the model in a human lymphoblastoid mutation assay. Mutat Res 216(3):149-61.

Pearson JV, Huentelman MJ, Halperin RF, Tembe WD, Melquist S, Homer N, Brun M, Szelinger S, Coon KD, Zismann VL and others. 2007. Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. Am J Hum Genet 80(1):126-39.

Potten CS, Owen G, Booth D. 2002. Intestinal stem cells protect their genome by selective segregation of template DNA strands. J Cell Sci 115(Pt 11):2381-8.

Rabbee N, Speed TP. 2006. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 22(1):7-12.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W and others. 2006. Global variation in copy number in the human genome. Nature 444(7118):444-54.

Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. 2001. Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. Nat Immunol 2(6):530-6.

Rubin H. 2006. What keeps cells in tissues behaving normally in the face of myriad mutations? Bioessays 28(5):515-24.

Shigenaga MK, Hagen TM, Ames BN. 1994. Oxidative damage and

mitochondrial decay in aging. Proc Natl Acad Sci U S A 91(23):10771-8.

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N and others. 2006. The consensus coding sequences of human breast and colorectal cancers. Science 314(5797):268-74.

Stoler DL, Chen N, Basik M, Kahlenberg MS, Rodriguez-Bigas MA, Petrelli NJ, Anderson GR. 1999. The onset and extent of genomic instability in sporadic colorectal tumor progression. Proc Natl Acad Sci U S A 96(26):15121-6.

Suh Y, Vijg J. 2006. Maintaining genetic integrity in aging: a zero sum game. Antioxid Redox Signal 8(3-4):559-71.

van den Hurk JA, Meij IC, Del Carmen Seleme M, Hoefsloot LH, Sistermans EA, de Wijs IJ, Plomp AS, de Jong PT, Kazazian HH, Cremers FP. 2007. L1 retrotransposition can occur early in human embryonic development. Hum Mol Genet.

Wang TL, Rago C, Silliman N, Ptak J, Markowitz S, Willson JK, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE. 2002. Prevalence of somatic alterations in the colorectal cancer cell genome. Proc Natl Acad Sci U S A 99(5):3076-80.

Weinberg HS, Korol AB, Kirzhner VM, Avivi A, Fahima T, Nevo E, Shapiro S, Rennert G, Piatak O, Stepanova EI and others. 2001. Very high mutation rate in offspring of Chernobyl accident liquidators. Proc Biol Sci 268(1471):1001-5.

Weiss KM. 2005. Cryptic causation of human disease: reading between the (germ) lines. Trends Genet 21(2):82-8.

Wong KK, Maser RS, Bachoo RM, Menon J, Carrasco DR, Gu Y, Alt FW, DePinho RA. 2003. Telomere dysfunction and Atm deficiency compromises organ homeostasis and accelerates ageing. Nature 421(6923):643-8.

Youssoufian H, Pyeritz RE. 2002. Mechanisms and consequences of somatic mosaicism in humans. Nat Rev Genet 3(10):748-58.