LETTER

doi:10.1038/nature11273

Chromatin organization is a major influence on regional mutation rates in human cancer cells

Benjamin Schuster-Böckler[1,2] & Ben Lehner[1,3]

Cancer genome sequencing provides the first direct information on how mutation rates vary across the human genome in somatic cells[1–7]. Testing diverse genetic and epigenetic features, here we show that mutation rates in cancer genomes are strikingly related to chromatin organization. Indeed, at the megabase scale, a single feature—levels of the heterochromatin-associated histone modification H3K9me3—can account for more than 40% of mutation-rate variation, and a combination of features can account for more than 55%. The strong association between mutation rates and chromatin organization is upheld in samples from different tissues and for different mutation types. This suggests that the arrangement of the genome into heterochromatin- and euchromatin-like domains is a dominant influence on regional mutation-rate variation in human somatic cells.

Comparative genomics and population studies suggest that human germline mutation rates are not constant across the genome[8,9]. Many genetic and epigenetic properties have been proposed to influence the rate of single nucleotide changes, including local base composition[8,10], DNA replication timing[8,11], chromatin structure[12] and the formation of double-strand breaks[13]. The sequencing of cancer genomes provides a unique opportunity to assess directly how mutation rates vary across the human genome[1–7]; by subtracting base changes observed in normal tissue from the same individual, a set of somatic single nucleotide variants (SNVs) can be derived. Moreover, the large number of genome-scale data sets available for human somatic cells enables the investigation of potential causes of mutation rate variation. It has been noted that tumours from different tissues show biases in mutation type[2,3] and evidence of transcription-coupled repair[3,6]. In addition, another study[7] showed that, at the 1-megabase (Mb) scale, there is substantial variation in the density of somatic mutations along the human genome and, moreover, that this regional variation in mutation density was correlated in three different tumour genomes. They also showed that somatic mutation rates measured in cancer genomes moderately correlate with those inferred in the germline from human–chimp sequence divergence[7] (Supplementary Fig. 1). However, so far, the individual features associated with mutation-rate variation explain very little of the regional variance across the genome[7].

We gathered a total of 84,879 unique SNV positions from individual leukaemia[5], melanoma[3], small cell lung cancer[2] and prostate cancer[4] genomes. To identify potential causes of mutation-rate variation across the genome, we compiled a set of diverse genetic and epigenetic features that have been mapped genome-wide in human cells. In total we examined 46 features, including base composition, CpG content, gene density, DNA replication timing[14], nucleosome occupancy[15], long-range chromatin interactions (Hi-C)[16], recombination rate[17], the density of unique sequences (mappability of 24-base polymers[18]), levels of 18 histone acetylations[19], levels of 17 histone methylations[20], and occupancy of RNA polymerase II, the CTCF insulator protein and the histone variant H2AZ[20]. We then calculated the correlation coefficient for all pairwise combinations of features, including cancer SNV density, human–chimp divergence and germline single nucleotide polymorphism (SNP) density, and clustered the features using these correlation coefficients as a distance metric.

Surprisingly, we found that at the megabase scale, cancer SNV density is strikingly correlated with many features of somatic cell chromatin organization (Fig. 1). The feature most strongly correlated with cancer SNV density is the repressive histone modification H3K9me3 ($r = 0.64$, $P < 2.2 \times 10^{-16}$). Positive correlations are also observed with other repressive marks including H3K9me2 ($r = 0.53$, $P < 2.2 \times 10^{-16}$) and H4K20me3 ($r = 0.39$, $P < 2.2 \times 10^{-16}$). In contrast, cancer SNV density anti-correlates with levels of many histone modifications associated with open chromatin, such as H3K4me3 ($r = -0.59$, $P < 2.2 \times 10^{-16}$) and H3K9ac ($r = -0.59$, $P < 2.2 \times 10^{-16}$). More moderate anti-correlation is observed with GC content ($r = -0.47$, $P < 2.2 \times 10^{-16}$), gene density ($r = -0.42$, $P < 2.2 \times 10^{-16}$), early replication timing ($r = -0.30$, $P < 2.2 \times 10^{-16}$) and the density of highly positioned nucleosomes ($r = -0.43$, $P < 2.2 \times 10^{-16}$). These conclusions are upheld when using alternative genomic window sizes (Fig. 1a): for example, the correlation with H3K9me3 is 0.37 at 100-kilobase resolution and 0.69 at 10-Mb resolution. We note that the weaker correlations when considering smaller window sizes may be due to the low median number of SNVs per window (Supplementary Fig. 10). Taken together this shows that, at least at the megabase scale, regional mutation-rate variation is strongly associated with regional variation in chromatin organization.

We used principal component analysis to investigate further the inter-dependencies among the different chromosome features. This revealed that at 1-Mb resolution, nearly 60% of the variance in these diverse features could be accounted for by a first principal component (Supplementary Fig. 3b). That is, the regional variation in many different genetic and epigenetic features can be captured by variation in a single underlying component along the genome. Features with a strong loading on this first principal component include many histone modifications and other features classically associated with either highly accessible euchromatin or inaccessible heterochromatin (Supplementary Fig. 3a). For example, the histone modifications H3K9me3 and H4K20me3 have strong negative loadings on this component, and GC content, gene density, early DNA replication and many activation marks show strong positive loadings (Supplementary Fig. 3a). Cancer SNV density also has a strong negative loading on this first component, consistent with the idea that somatic mutation rates in cancer cells are highest in inaccessible, heterochromatin-like regions and lowest in accessible euchromatin-like domains. In contrast, germline SNP density and human–chimp divergence have stronger loadings on the second orthogonal principal component (Supplementary Fig. 3). Indeed, consistent with previous findings[13] and an important role for background selection and/or genetic hitchhiking in determining human diversity levels[21], germline SNP density is most positively correlated with the rate of recombination during meiosis ($r = 0.45$, $P < 2.2 \times 10^{-16}$).

To confirm that our conclusions were not tumour-type-specific, we also analysed the mutations from each cancer sample in isolation. The

[1]EMBL-CRG Systems Biology Unit, CRG and UPF, Barcelona 08003, Spain. [2]Pear Computer LLP, London W5 1SH, UK. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys, 23, Barcelona 08010, Spain.

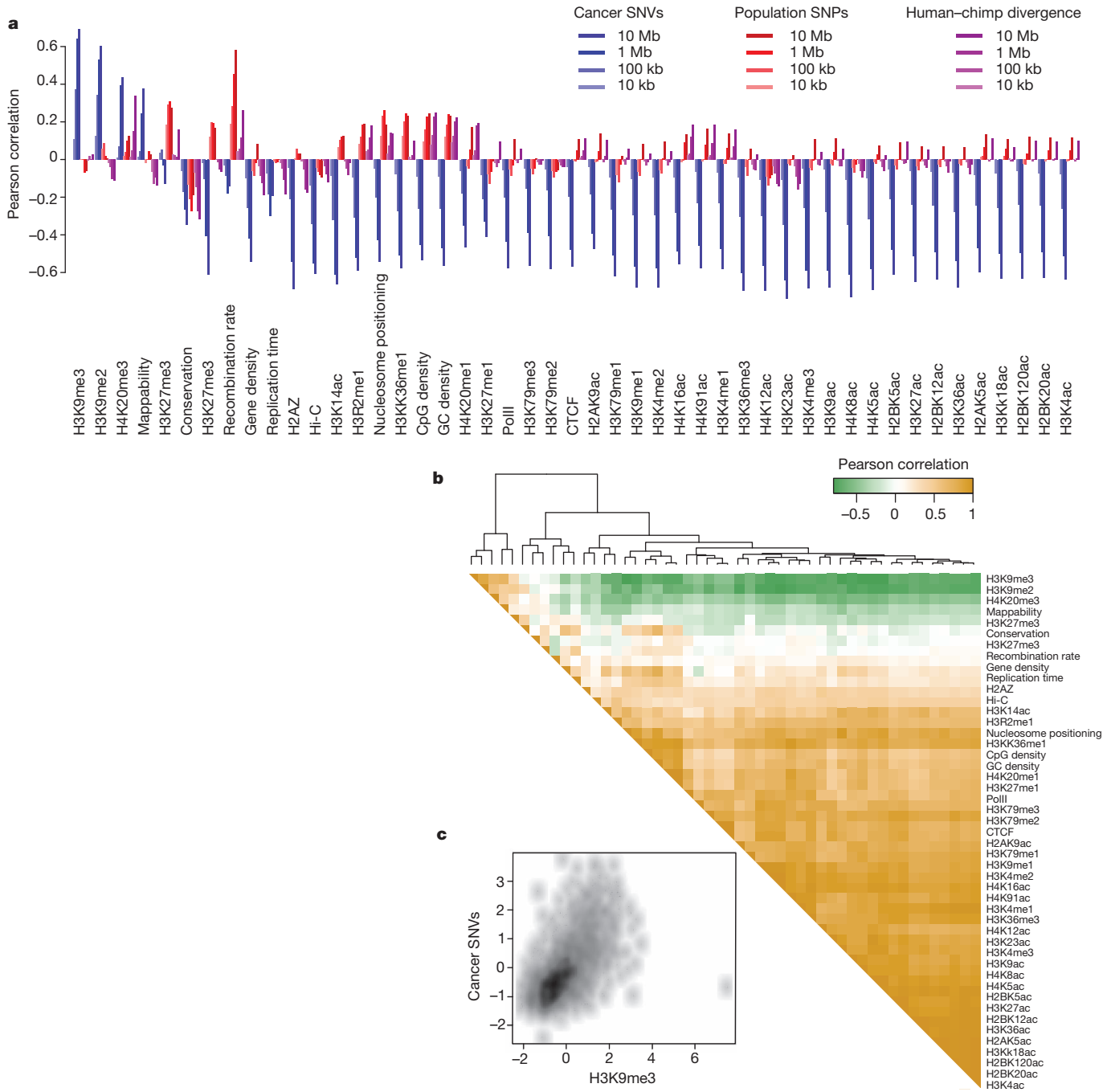

Figure 1 | The density of somatic mutations in cancer genomes correlates with H3K9me3 modification levels and anti-correlates with genomic features associated with open chromatin. **a**, Pearson correlation coefficients of cancer SNVs (blue), dbSNP density (red) and human–chimp divergence (purple) with genomic features in non-overlapping, non-repetitive windows of different sizes along the genome. **b**, The correlation matrix. Dark green denotes negative and yellow positive correlation at 1-Mb resolution. **c**, Smoothed scatter-plot of cancer SNV density against H3K9me3 modification levels, both normalized to standard scores. Correlation plots for all other features are available in Supplementary Figs 5–8. The measure of replication timing used here is high for early replicating regions.

individual samples derived from distinct tissues and showed signatures of exposure to different environmental mutagens such as ultraviolet radiation in the melanoma[3] and carcinogens from tobacco smoke in the lung cancer sample[2]. However, SNV density is positively correlated with levels of H3K9me3 and negatively correlated with many features associated with open chromatin in each of the individual tumour samples, supporting the generality of our findings (Fig. 2).

Considering transition mutations separately from transversions, or CpG mutations separately from non-CpG mutations, also does not change our conclusions: elevated mutation rates are strongly associated with H3K9me3 (Fig. 3a) and other indicators of heterochromatin (Supplementary Fig. 11), irrespective of the mutation type. Likewise, the association remains strong when considering only non-genic or only genic regions of the genome (Fig. 3b), so cannot be accounted for by transcription- or expression-coupled repair. The correlation is also strong when only considering SNVs surrounded by unique sequence (Fig. 3c and Supplementary Fig. 9), after filtering out regions of the genome with extreme GC content (Fig. 3c), and when excluding evolutionarily conserved bases (Fig. 3c). The association between chromatin organization and mutation-rate variation is therefore upheld for diverse tissue types, diverse mutation types and diverse genomic regions.

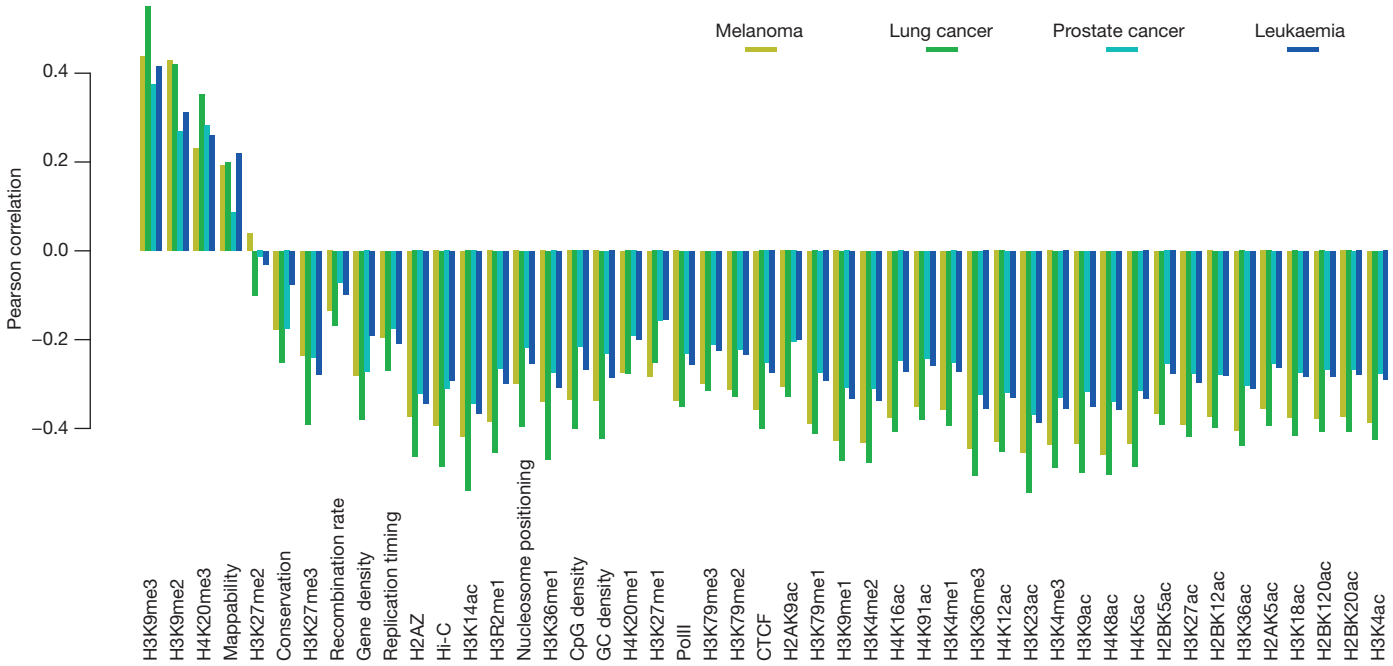

Figure 2 | Correlation coefficients of SNV density from individual cancer genomes at 1-Mb resolution with diverse genetic and epigenetic features.

Last, we examined the extent to which predictions of mutation-rate variation could be improved by combining the information from multiple genomic features. We used an iterative procedure to identify the combination of features that provide the best predictions in multiple linear regression models using increasing numbers of features (see Methods) and found that more than 55% of the variance in cancer SNV density along the genome could be explained by combining features (Fig. 4). In contrast, predictions of germline SNP density or human–chimp divergence never accounted for more than 35% of the variance, with recombination rate alone accounting for 20.5% of the observed variance in germline SNP density (Supplementary Fig. 4). In the cancer cells, H3K9me3 alone can account for more than 40% of the observed variance in SNV density. The remaining predictive features included in the models mark regions with open chromatin, or in the case of Hi-C, further distinguish the compartmental organization of the genome. The Hi-C metric used here is a measure devised previously[16] and uses genome-wide data on physical contacts between regions through three-dimensional folding of the chromosome. It distinguishes between densely packed chromatin with strong short-range interactions and accessible euchromatin with a more diverse interaction pattern. The Hi-C metric anti-correlates with somatic SNV density ($r = -0.55, P < 2.2 \times 10^{-16}$), further supporting a model in which chromatin organization is a major determinant of variation in regional mutation rate.

The epigenetic modifications analysed here were not profiled in the same cell types as the somatic mutations, which suggests that the actual influence of chromatin organization on regional mutation rates may have been underestimated in our analyses. Furthermore, chromatin could also have an important influence on germline mutation rates, particularly if chromatin organization in the germline is substantially different to that in somatic cells[22]. An improved understanding of chromosome organization in the germline will be required to test this possibility.

However, at least in cancer cells, our analyses indicate that the dominant determinant of regional mutation rate variation is chromatin

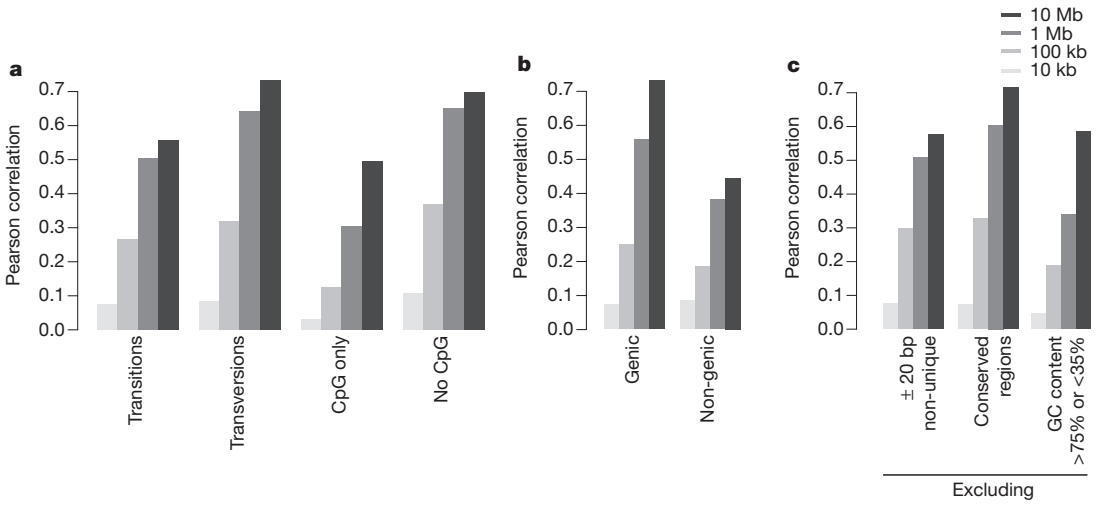

Figure 3 | Chromatin organization predicts cancer genome mutation density for diverse mutation types and sequence contexts. **a**–**c**, Correlations by mutation type (**a**), in genic or non-genic regions only (**b**), and only considering SNVs surrounded by 20 base pairs of unique sequence, after filtering out regions of the genome with average GC density of less than 35% or more than 75%, or when excluding evolutionarily conserved bases (**c**) (see Methods).

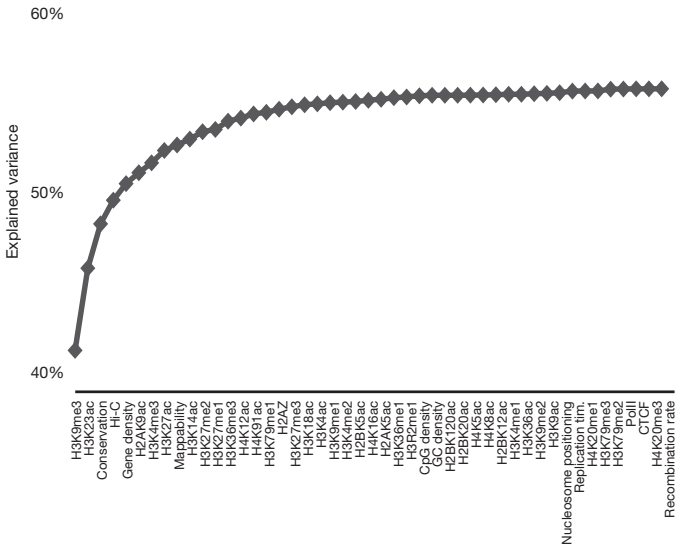

Figure 4 | Prediction of cancer SNV density variation using integrated models. Cumulative R^2 of linear models, adding the feature on the x axis as a predictor at each step.

organization, with mutation rates elevated in more heterochromatin-like domains and repressed in more open chromatin. This could reflect differing accessibility to DNA repair complexes[23], variation in the ability to signal repair[24] or perhaps increased exposure to mutagens at the nuclear periphery[25]. The somatic mutations considered here all arose in lineages that ultimately gave rise to tumours; although the mutation process may be different in tumour lineages, the tumours analysed here derive from diverse tissue types, which suggests the intriguing possibility that chromatin organization will be a major influence on regional mutation rates in all human somatic cells.

METHODS SUMMARY

SNVs of human cancer cells were obtained from recent publications[1–5]. Germline polymorphisms were taken from dbSNP[26] and the 1000 genomes project[27]. Ensembl Compara provided the human–chimp divergence data[28]. Histone methylation[20] and acetylation[19] states were mapped to the genome, as well as an array of additional genomic feature sets from various sources: recombination rate[17], nucleosome positioning[15,29], spatial proximity[16], replication timing[14], gene density and evolutionary conservation[30]. Genomes were then split into evenly sized windows, and windows with a high repeat content[18] were excluded to calculate Pearson correlations between features.

Full Methods and any associated references are available in the online version of the paper.

Received 29 September 2011; accepted 31 May 2012.
Published online 22 July 2012.

1. Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456,** 66–72 (2008).
2. Pleasance, E. D. *et al.* A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463,** 184–190 (2010).
3. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191–196 (2010).
4. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470,** 214–220 (2011).
5. Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475,** 101–105 (2011).
6. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465,** 473–477 (2010).
7. Hodgkinson, A., Chen, Y. & Eyre-Walker, A. The large-scale distribution of somatic mutations in cancer genomes. *Hum. Mutat.* **33,** 136–143 (2012).
8. Wolfe, K. H., Sharp, P. M. & Li, W. H. Mutation rates differ among regions of the mammalian genome. *Nature* **337,** 283–285 (1989).
9. Ellegren, H., Smith, N. G. C. & Webster, M. T. Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.* **13,** 562–568 (2003).
10. Cooper, D. N. & Krawczak, M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Hum. Genet.* **83,** 181–188 (1989).
11. Stamatoyannopoulos, J. *et al.* Human mutation rate associated with DNA replication timing. *Nature Genet.* **41,** 393–395 (2009).
12. Prendergast, J. G. D. *et al.* Chromatin structure and evolution in the human genome. *BMC Evol. Biol.* **7,** 72 (2007).
13. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18,** 337–340 (2002).
14. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci. USA* **107,** 139–144 (2010).
15. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).
16. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
17. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
18. Dreszer, T. R. *et al.* The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* **40,** D918–D923 (2012).
19. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genet.* **40,** 897–903 (2008).
20. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
21. Gossmann, T. I., Woolfit, M. & Eyre-Walker, A. Quantifying the variation in the effective population size within a genome. *Genetics* **189,** 1389–1402 (2011).
22. Vavouri, T. & Lehner, B. Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome. *PLoS Genet.* **7,** e1002036 (2011).
23. Peterson, C. L. & Côté, J. Cellular machineries for chromosomal DNA repair. *Genes Dev.* **18,** 602–616 (2004).
24. Goodarzi, A. A. *et al.* ATM signaling facilitates repair of DNA double-strand breaks associated with heterochromatin. *Mol. Cell* **31,** 167–177 (2008).
25. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128,** 787–800 (2007).
26. Wheeler, D. L. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **35,** D5–D12 (2007).
27. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
28. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18,** 1814–1828 (2008).
29. Zhang, Y., Shin, H., Song, J. S., Lei, Y. & Liu, X. S. Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC Genomics* **9,** 537 (2008).
30. Broad Institute Sequencing Platform and Whole Genome Assembly Team *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–481 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was funded by an European Research Council (ERC) Starting Grant, European Union Framework 7 project 277899 4DCellFate, ERASysBioPLUS, Ministerio de Ciencia e Innovación (MICINN) grants BFU2008-00365 and BFU2011-26206, Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), the European Molecular Biology Organization (EMBO) Young Investigator Program, the EMBL-CRG Systems Biology Program and a Juan de la Cierva postdoctoral fellowship to B.S-B. We thank T. Vavouri and T. Warnecke for comments on the manuscript, and R.S. Hansen for assistance with analysing replication timing data.

Author Contributions B.S.-B. performed all analyses. B.S.-B. and B.L. designed analyses and wrote the manuscript. B.L. conceived the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.L. (ben.lehner@crg.es).

METHODS

Cancer single nucleotide variation data. Autosomal SNVs of human cancer cells were obtained from the supplementary data files of the respective publications: 32,075 SNVs for melanoma[3], 27,246 for prostate cancer[4], 21,707 for lung cancer[2] and 3,861 SNVs for leukaemia[5]. For the leukaemia data, only the locations of mutations were available, so they are not included in the calculation of transition/transversion correlations.

Germline polymorphism data. dbSNP build 130 comprising 4,118,806 SNPs was downloaded from the NCBI FTP server[26]. SNP data from 1000 genomes pilot 2 for two family trios of Central European and West African descent were downloaded from the 1000 genomes FTP server, yielding 5,739,704 unique SNPs[27]. Where necessary, coordinates were mapped to the NCBI36/hg18 assembly of the human genome using the liftOver tool with standard parameters[18].

Human–chimp sequence divergence. Divergence data between *Homo sapiens* and *Pan troglodytes* were extracted from EPO (Enredo–Pecan–Ortheus) whole-genome alignments[28] available from Ensembl release 54. The species tree used to construct the alignment contained genomic sequences of human, chimp, orangutan, macaque, mouse, rat, dog, horse and cow. This enabled the inference of ancestral alleles for each change. The human–chimp alignment covers more than 88% of the human genome, yielding approximately 10^8 substitutions across all autosomes.

Genome-wide feature sets. Recombination rates were reported using data from the deCODE genetic map[17] at 1-Mb resolution. The genome-wide uniqueness of 24-base polymers was calculated by the ENCODE project and downloaded from the University of California, Santa Cruz (UCSC) genome browser[18]. Highly positioned nucleosomes were predicted with the NPS algorithm[29] using short read densities from micrococcal nuclease digested chromatin extracted from resting CD4 T cells as reported by Schones *et al.*[15]. Short sequence tags for histone methylations, H2AZ, CTCF, PolII binding[20] and histone acetylations[19] were converted to read densities by sliding a 160-base-pair window across each chromosome, summing up reads 80 base pairs upstream on the forward and 80 base pairs downstream on the reverse strand[15]. Hi-C eigenvectors at 100-kilobase resolution for the lymphoblastoid cell line GM06990 were downloaded from the Gene Expression Omnibus (GEO) entry with accession number GSE18199. In accordance with the original paper, eigenvector 1 was used for all chromosomes except chromosomes 4 and 5 for which eigenvector 2 was chosen. Replication timing was calculated using data from Hansen *et al.*[14]. Raw reads were downloaded for four cell-cycle fractions (G1B, S1, S4 and G2), mapped to the NCBI36/hg18 assembly of the human genome using the GEM library (http://big.crg.cat/services/gem_genome_multi_tool_library), averaged to 4 million reads each, and normalized to percentage replication per nucleotide position. Finally, an early-to-late ratio was calculated as $(G1B + S1)/(S4 + G2)$. CpG density is the fraction of residues in CpG dinucleotides. GC density refers to the fraction of all G or C residues per window. Gene density refers to the fraction of nucleotides covered by a gene (including introns) per window. Locations of known coding genes were downloaded from EnsEMBL BioMart (release 56) and mapped to the NCBI36/hg18 assembly of the human genome using the liftOver tool[18]. Evolutionarily conserved bases were identified using the recently published analysis of 29 mammalian genomes[30].

Filtering of data. The mappability feature described above assigns values of 1 to unique 24-base polymers in the genome, 0.5 to those that occur twice, 0.33 to those that occur three times, and 0 otherwise. We empirically established a conservative cut-off value of 0.8 for the average mappability per window, excluding windows with highly repetitive DNA elements. We used the log-odds Siphy-π scores with a cut-off threshold of 3 to identify conserved residues. To mask certain genomic regions while calculating the window averages of each property, we only counted those nucleotides in each genomic window that were not masked by the applied filter. Individual windows were excluded from the analysis if more than 90% of the nucleotides were masked.

Principal component analysis. Principal component analysis was performed on the mappability-filtered matrix of 2015 rows representing 1-Mb windows and columns corresponding to 47 genomic features as well as cancer SNV, germline SNP and human–chimp divergence densities. Calculations were performed in R using the princomp function. All feature vectors were scaled to mean 0 and standard deviation 1.

Iterative model refinement. To identify the minimal informative set of predictive features for somatic SNV, germline SNP and human–chimp divergence densities, linear models were fitted by generalized least-squares estimation between each individual feature and somatic/germline SNP density. We compared the models by their Akaike information criterion (AIC) and chose the feature with minimal AIC. This procedure was repeated 46 times, adding one feature to the model at each iteration. The set of features with minimal AIC was chosen as the minimal informative set of predictive features. Percentage explained variance was calculated as the R^2 of a linear regression model using these sets of predictive features. Calculations were performed in R using the AIC, gls and lm functions.