



# Cell-of-origin chromatin organization shapes the mutational landscape of cancer

## Citation

Polak, P., R. Karlić, A. Koren, R. Thurman, R. Sandstrom, M. Lawrence, A. Reynolds, et al. 2015. "Cell-of-origin chromatin organization shapes the mutational landscape of cancer." *Nature* 518 (7539): 360-364. doi:10.1038/nature14221. <http://dx.doi.org/10.1038/nature14221>.

## Published Version

doi:10.1038/nature14221

## Permanent link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:21462011>

## Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#).

[Accessibility](#)



Published in final edited form as:

*Nature*. 2015 February 19; 518(7539): 360–364. doi:10.1038/nature14221.

## Cell-of-origin chromatin organization shapes the mutational landscape of cancer

**Paz Polak**<sup>#1,2</sup>, **Rosa Karlič**<sup>#3</sup>, **Amnon Koren**<sup>4,2</sup>, **Robert Thurman**<sup>5</sup>, **Richard Sandstrom**<sup>5</sup>, **Michael Lawrence**<sup>2</sup>, **Alex Reynolds**<sup>5</sup>, **Eric Rynes**<sup>5</sup>, **Kristian Vlahoviček**<sup>3,6</sup>, **John A. Stamatoyannopoulos**<sup>5,#</sup>, and **Shamil R. Sunyaev**<sup>1,2,#</sup>

<sup>1</sup> Division of Genetics, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA, 02115

<sup>2</sup> The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.

<sup>3</sup> Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia

<sup>4</sup> Department of Genetics, Harvard Medical School, Boston, MA 02115, USA.

<sup>5</sup> Departments of Genome Sciences and Medicine (Oncology), University of Washington, Seattle, WA 98195, USA

<sup>6</sup> Department of Informatics, University of Oslo, P.O. Box 1080, Blindern, NO-0316 Oslo, Norway

# These authors contributed equally to this work.

### Abstract

Cancer is a disease potentiated by mutations in somatic cells. Cancer mutations are not distributed uniformly along the genome. Instead, different genomic regions vary by up to 5-fold in the local density of somatic mutations<sup>1</sup>, posing a fundamental problem for statistical methods of cancer genomics. Epigenomic organization has been proposed as a major determinant of the cancer mutational landscape<sup>1-5</sup>. However, both somatic mutagenesis and epigenomic features are highly cell-type-specific<sup>6,7</sup>. We investigated the distribution of mutations in multiple samples of diverse cancer types and compared them to cell-type-specific epigenomic features. Here, we show that chromatin accessibility and modification, together with replication timing, explain up to 86% of the variance in mutation rates along cancer genomes. Overwhelmingly, the best predictors of local somatic mutation density are epigenomic features derived from the most likely cell type of origin of the corresponding malignancy. Moreover, we find that cell-of-origin chromatin features are much stronger determinants of cancer mutation profiles than chromatin features of cognate cancer cell lines. We show further that the cell type of origin of a cancer can be accurately determined based on the distribution of mutations along its genome. Thus, DNA sequence of a cancer genome encompasses a wealth of information about the identity and epigenomic features of its cell of origin.

#Correspondence: Shamil Sunyaev <ssunyaev@rics.bwh.harvard.edu>, John Stamatoyannopoulos <jstam@u.washington.edu>.

Author Contributions

S.S., J.A.S., P.P. and R.K. conceived the project and provided leadership. P.P, R.K., A.K., M.S.L., R.S., R.T A.S., A.R and K.V. analyzed the data and contributed to scientific discussions. S.S, P.P., R.K., A.K and J.A.S. wrote the paper.

Recent studies have begun to address the underlying causes of cancer mutational heterogeneity by comparing mutation rate variation to the distribution of sequence features, gene expression and epigenetic marks along the genome<sup>2-5</sup>. A major limitation of previous studies was their uniform treatment of mutations from different cancers, and their consideration of epigenetic marks from a single cell type, usually a cell type different from the cancer tissue of origin. However, cancer is far from being a disease of uniform origin, progression and cell biology. Instead, different cancer types differ in their overall mutation rates, their predominant mutation types, and the distribution of mutations along their genomes<sup>1</sup>. Substantial variation also exists in the epigenomic landscape of different tissues, specifically in patterns of chromatin accessibility, histone modifications<sup>8</sup> [EC00], gene expression and DNA replication timing<sup>9</sup>. The full understanding of the factors contributing to mutational heterogeneity in cancer genomes thus requires the evaluation of the relationship between multiple epigenetic marks and mutation patterns in a cell-type-specific manner.

We analyzed a total of 173 cancer genomes from eight different cancer types that represent a wide range of tissues of origin, carcinogenic mechanisms, and mutational signatures: melanoma<sup>10</sup>, multiple myeloma<sup>11</sup>, lung adenocarcinoma<sup>12</sup>, liver cancer<sup>13</sup>, colorectal cancer<sup>14</sup>, glioblastoma<sup>15</sup>, esophageal adenocarcinoma<sup>16</sup>, and lung squamous cell carcinoma<sup>17</sup>. Regional variations in mutation density appeared similar although not identical among the different cancer types (Extended Data Fig. 1).

We compared the genomic distribution of mutations in these cancer genomes to 424 epigenetic features that were measured by the Epigenome Roadmap consortium [EC00]. These features were derived from 106 different cell types from 45 different tissue types, including the cell types of origin of most of the cancer types that we investigated (Methods and Extended Data Fig. 2). Importantly, the data represent primary human cells rather than cell lines. These epigenetic features comprised eight different types of variables, including DNaseI hypersensitive sites (DHS) (a global measure of chromatin accessibility)<sup>7</sup> and various histone modifications. An example of the variation in mutation density along chromosomes at 1Mb scale together with a representative epigenetic mark (DHS) is shown in Figure 1. In this case, as in most other cases (see below), epigenetic marks indicative of open chromatin and high gene activity were associated with low mutation density, while repressive, closed chromatin marks were associated with regions of high mutation density. Notably, these statistical associations do not necessarily imply causal effects of individual chromatin features, nor point to specific biological mechanisms.

The comparison of individual epigenomic features with local mutation density revealed that chromatin marks corresponding to the tumor's cell type of origin are more strongly associated with local mutation density than marks corresponding to unrelated cell types. For example, DHS marks from melanocytes explained a substantially larger fraction of the variance in melanoma mutation density than DHS marks from other cell types, even from the same tissue (skin) (Figure 1b). As another example, even though H3K4me1 marks in melanocytes and hepatocytes are highly correlated ( $r=0.8$ ), the distribution of mutations in liver cancer followed the levels of H3K4me1 in hepatocytes, but not in melanocytes, while

melanoma mutations correlated with the levels of H3K4me1 in melanocytes but not in hepatocytes (Figure 1c).

This initial observation suggested that the impact of chromatin on local mutation density is highly cell-type-specific. The comprehensive representation of different cell types in the Epigenome Roadmap could thus enable an improved prediction accuracy of mutations compared to previous studies. To rigorously quantify the contribution of different chromatin marks and gene expression to regional mutation density, and the extent of cell type specificity, we used Random Forest regression (Methods).

Remarkably epigenetic marks, together with replication timing measured in ENCODE cell lines<sup>18</sup>, collectively explained 74-86% of the variance in mutation density in seven cancer types (Figure 2a). In glioblastoma, for which fewer mutations were available for the analysis, 55% of the variance in mutation density could be explained. This is substantially higher than in earlier studies<sup>4</sup> and indicates that, at least for these cancer types, we have identified a set of epigenetic variables and cell types that almost fully predict the mutational variability along the genome. This enhanced prediction accuracy was not simply due to the larger size of the training data relative to previous studies, as the predictive ability dropped by only ~2-6% when only 10% of the data was used (Extended Data Fig. 3).

Prediction accuracy in individual samples is expected to be lower than in samples pooled by cancer type due to tumor heterogeneity, sampling variance, and a lower number of mutations available for the analysis (Extended Data Fig. 4). To evaluate the influences of these variables on the prediction ability of the Random Forest model, we simulated mutation datasets of variable sizes generated by the model itself, and compared the prediction accuracy of simulated and real data as a function of the number of mutations. For most samples, epigenomic features explained most (on average 70%) of the maximally predicted variance (Extended Data Fig. 5), and more than was explained by earlier studies<sup>2,4</sup> when matching dataset sizes. As a point of direct comparison with an earlier study<sup>2</sup> that did not use cell type specific chromatin marks, our model explained 50% of the variance in mutation density in the melanoma cell line COLO829<sup>19</sup>, for which the earlier study explained 29% of the variance.

The prediction accuracy was similar whether testing for all mutations or only the mutations of the predominant type (Extended Data Fig. 6) in each cancer type<sup>1,20</sup> (Figure 2b). A notable exception was lung adenocarcinoma, where a larger fraction of the variance could be explained for G>T mutations associated with smoking<sup>1,12,21</sup> than for C>T mutations. This difference was observed for both samples with G>T transversions<sup>12</sup> and C>T transitions as the leading mutational sources (Figure 2c).

Interestingly, prediction accuracy was fully explained by chromatin features, with gene expression and nucleotide content not providing any further improvement to the accuracy of the model. Even though gene expression has been unequivocally demonstrated to influence mutation density, chromatin features appear to be statistically stronger predictors (Extended Data Fig. 7).

When considering individual contributions to mutation rate prediction, between six and nineteen variables passed the significance threshold in any individual cancer type. There was a sweeping association between cancer mutations and chromatin marks measured in the cell type of origin of each cancer (Figures 3a). For instance, six out of the top ten features explaining variation in melanoma mutation density were derived from melanocytes (Figure 1 and Figure 3b). Similarly, seven out of the nine top features explaining mutational profiles in liver cancer were measured in liver cells. Comparable results were obtained for multiple myeloma, colorectal adenocarcinoma and glioblastoma, where most of the significant features were measured in hematopoietic, intestine mucosa and brain tissues, respectively. For esophageal adenocarcinoma, the top predictors were chromatin features derived from stomach mucosa rather than from esophageal tissues; this is expected given that the analyzed esophageal adenocarcinomas were triggered by Barrett's esophagus cells that resemble stomach epithelial cells<sup>22</sup>. Lung adenocarcinoma and lung squamous cell carcinoma were the only exceptions in that the top predictors were scattered among different tissue groups; the lack of tissue specificity in these cases likely results from the absence of epigenetic marks from normal lung epithelial cells in our dataset.

The results of the Random Forest regression were confirmed using backward feature selection to identify the minimal set of epigenetic predictors of mutations in each cancer type (Methods). As few as three to five features were sufficient to capture the variance explained by the full set of 424 different features (Extended Data Fig. 8), and in all cancers besides lung (as above), most of these features were derived from the corresponding cell types of origin. As a more direct test, we grouped all epigenomic data by cell or tissue type and compared the collective explanatory power of chromatin features derived from the cell types of origin vs. unmatched cell types. The results of this analysis confirmed the cell type specificity of the association between chromatin features and mutation density (Figure 3c).

The above results pose a key question on whether epigenomic features derived from the cell type of origin are the strongest determinants of cancer mutations, or whether they simply serve as the best available proxies to the chromatin organization of the corresponding malignant cells. The availability of epigenomic data for the liver cancer cell line HepG2<sup>8</sup> and for melanoma cell lines made it possible to directly address this question. Surprisingly, in both cases, epigenomic features from the cell type of origin resulted in a higher prediction accuracy than those from the cancer cell lines. The Random Forest predictor trained on chromatin features of HepG2 was less accurate in predicting the liver cancer mutation density than the analogous predictor trained on features of hepatocytes (Figure 3d). Similarly, chromatin accessibility in melanocytes was a much better predictor of mutation density in the COLO829 melanoma cell line (Figure 3e and Extended Data Fig. 9). Thus, chromatin features associated with carcinogenesis do not determine cancer mutations to the same extent as chromatin features of the cells of origin. We envision two potential explanations for this observation. First, most of the mutations observed in cancers may arise prior to the epigenetic changes linked to neoplastic progression. In addition, advanced tumors may undergo specific epigenetic changes that distinguish them from other tumors of the same type.

Taken together, the above results strongly suggest that the cell of origin of an individual tumor sample could be predicted from its mutation pattern alone. Mutation profiles of individual samples cluster according to cancer type, and, consequently cell of origin (Figure 4a). We developed a straightforward predictor based on enrichment of epigenomic variables from a single cell type among the top 20 variables selected by the Random Forest analysis. This approach classified 88% of melanoma, colorectal, liver, multiple myeloma, esophageal and glioblastoma cancer genomes to melanocytes, colonic mucosa, liver, hematopoietic, stomach mucosa and brain tissues, respectively (Figure 4b). Thus, mutational patterns contain sufficient information for identifying the cell type of origin of a tumor. We propose that sequencing the DNA of a tumor of unknown primary origin can allow the precise pinpointing of the cell type of origin of that tumor.

Traditionally, statistical prediction in cancer has made use of gene expression data. We therefore constructed an analogous predictor of cell of origin using RNA sequencing data from 167 glioblastoma multiforme and 370 skin cutaneous melanoma samples<sup>23</sup>. This predictor achieved accuracies of 78% and 57% on these cancer types, slightly lower than the mutation-based predictor. Although these two classifiers are not directly comparable, it is clear that genome sequence carries at least the same amount of information about the cell of origin as gene expression data does.

In conclusion, our observations suggest that cancer mutation density is linked to the epigenomic profile in a highly cell-type-specific manner. Thus, DNA sequence is informative about the origin of an individual tumor. The accumulating epigenomic data on human cell types opens the perspective for accurate prediction of the cell of origin of a cancer from its genome sequence.

## Methods

### Data

We divided the human genome into 1 MB regions, excluding regions overlapping centromeres and telomeres, as well as regions where the fraction of uniquely mappable base pairs was lower than 0.92. We calculated the mean signal for different histone modifications, DNase I hypersensitivity and replication timing in different cell types, and used these 424 features to predict mutation density along the genome in eight different cancer types (see below).

We calculated mutation density by obtaining data for 173 individual cancer genomes, belonging to eight cancer types: melanoma (25 genomes)<sup>10</sup>, lung adenocarcinoma (24 genomes)<sup>12</sup>, lung squamous cell carcinoma (12 genomes)<sup>17</sup>, esophageal adenocarcinoma (9 genomes)<sup>16</sup>, liver (64 genomes)<sup>13</sup>, multiple myeloma (23 genomes)<sup>11</sup>, colorectal cancer<sup>14</sup> (CRC, 9 genomes) and glioblastoma (7 genomes)<sup>15</sup>. The whole genome of the COLO-829 cell line has been sequenced by the Sanger Institute. The COLO-829 cell line was derived from metastatic tissue. The liver cancers were sequenced by the National Cancer Center Research Institute in Japan. The mutation lists for the COLO829 cell line and liver cancer that we used in this study can be found at [http://dcc.icgc.org/download/legacy\\_data\\_releases/version\\_07/](http://dcc.icgc.org/download/legacy_data_releases/version_07/) under the folders Malignant\_Melanoma-WTSI-UK

(COLO-829) and Liver\_Cancer-NCC-JP. The rest of the genomes were sequenced and analyzed by the Broad Institute and called using MuTect<sup>24</sup> (<http://www.broadinstitute.org/cancer/cga/mutect>).

For each cancer type we counted the overall number of mutations in all individual cancer genomes belonging to that cancer type. We also determined the mutation densities for all possible types of mutations in each cancer types by counting different types of mutations in 1 Mb windows and normalizing for the sequence composition of each window.

We downloaded data for 7 different histone modifications and DNase I hypersensitivity from Epigenomics Roadmap [EC00] and ENCODE<sup>8</sup> (Extended data Fig. 1). Epigenomic data is available from the NCBI via the GEO series GSE18927 for University of Washington Human Reference Epigenome Mapping Project at <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18927>. Data used in this study can also be viewed via multiple browsers outlined at the <http://roadmapepigenomics.org/> website.

Fetal tissues were obtained from morphologically normal fetuses by the Birth Defects Research Laboratory in the Department of Pediatrics at the University of Washington, collected under an IRB-approved protocol. Blood cell subsets were collected from fully consented, normal donors at the Cellular Therapy Laboratory and cGMP Cell Processing Facility under the direction of Shelly Heimfeld at the Fred Hutchinson Cancer Research Center with IRB-approval.

For histone modifications we combined reads for all samples belonging to one cell type and calculated RPKM values for 1 Mb windows along the genome. We also calculated the average number of DNase I hypersensitivity peaks overlapping 1 Mb windows across all samples belonging to a certain cell type. We used BEDOPS<sup>25</sup> to map reads and DHS peaks to intervals.

We obtained data for four different Repli-seq experiments from the ENCODE project (Extended Data Fig. 1) and determined replication timing as the average value of wavelet-smoothed signal in each 1 Mb window. Lymphoblastoid cell line replication time was obtained from Koren et al., 2012<sup>26</sup> and averaged over 1Mb windows along the genome.

To control for the effect of sequence features on mutation density, for each 1 Mb window we also calculated GC content, the number of CpG, GpC, and ApT dinucleotides, and fraction of the window overlapping coding regions, known genes and CpG islands.

To control for the effect of expression on mutation density we downloaded mRNA-seq data from the Epigenomics Roadmap [EC00], for 38 different cell types for which expression data was available (Extended Data Table 1). We combined reads for all samples belonging to one cell type and calculated RPKM values for the set of all protein coding exons in 1 Mb windows, the set of all protein coding and lncRNA exons in 1 Mb windows, the maximally expressed gene in a 1 Mb window or non-genic regions in 1 Mb windows.

## Random Forest regression

Random Forest is a non-parametric machine learning method that combines the output of an ensemble of regression trees to predict the value of a continuous response variable<sup>27</sup>. The use of multiple regression trees reduces the risk of over-fitting and makes the method robust to outliers and noise in the input data. For each regression tree, a training set of  $N$  samples are drawn, with replacement, from the dataset. The remaining data (out-of-bag data) constitutes the test set for this tree, and is used to compute the mean squared prediction error of the tree. The prediction for each sample is made by taking the average of predictions over all trees for which the sample was part of the out-of-bag data.

Random Forest provides an internal measure of the importance of different predictor variables, based on out-of-bag data. The mean squared error calculated on the out-of-bag data is recorded in every tree grown in the forest. The values of all the predictor variables are then randomly permuted in all the out-of-bag samples and the mean squared error is computed again. The difference between the two errors is averaged over all the trees, and normalized by the standard error, representing the raw importance score for each variable.

We used Random Forest with 1000 trees to predict mutation densities in 1Mb non-overlapping windows in the eight different cancer types using 424 predictor variables (epigenetic features and replication timing; Extended Data Fig. 1). We divided the data into ten non-overlapping sets and predicted the number of mutations in each cancer type using 10-fold cross-validation. For each sample, the predicted value corresponded to the predicted mutation density when this sample was part of the test set. We used Pearson product-moment correlation to interpret the prediction accuracy. The fraction of variance explained by each model was calculated as the Pearson correlation coefficient squared.

## Controlling for the effect of sequence features and expression on prediction accuracy

We created different subsets of features corresponding to chromatin (histone modifications and DNase hypersensitivity, 419 features), replication timing (5 features), sequence (7 features) and expression (38 features). We then used Random Forest regression with 10-fold cross-validation to predict mutation density in different cancers, where for each cancer type we trained different models: on each subset of features separately and on combinations of different subsets of features.

## Variable importance analysis

Variable importance was calculated for each predictor variable in each cancer type by permuting the variable, i.e. randomly shuffling the data values so that the relationship between the response and predictor variables was destroyed. The percent of increase in mean squared error of prediction was then calculated. Since the variable importance can be influenced by both the correlation and the scale of the variables, we calculated the empirical p-value of variable importance measures by repeatedly permuting the response variable in Random Forest models, in order to determine the distribution of measured importance values for each predictor variable<sup>28</sup>. This procedure was repeated 1000 times, and the number of times in which the importance measure in the original data set was lower or equal



to the permuted importance measure was counted; this count represented the p-value, with a count of one corresponding to a significance level of  $P < 0.001$ .

### **Feature selection**

We applied backward elimination to identify a minimal set of predictors for each cancer type. Backward elimination is a “greedy” algorithm which finds the locally optimal subset of features, but does not guarantee finding the global optimum. However, it is less computationally intensive than searching all possible feature subsets when the number of features (N) is large (in our case  $N=424$ ). Initially, we trained a Random Forest with 10-fold cross-validation on the complete set of variables and determined the importance of all the variables in the model (the importance was calculated as the mean importance of the variable across 10 rounds of cross-validation). We then ranked the variables according to their importance and determined the top 20 variables. We then sequentially trained 20 models, removing the least important variable at each step, until only one predictor variable was left for training.

### **Principal coordinate analysis**

We used principal coordinate analysis to visualize the dissimilarities in mutation density distributions between individual cancer genomes. Dissimilarity was calculated as  $1 -$  Pearson correlation coefficient, for all possible combinations of individual cancer genomes.

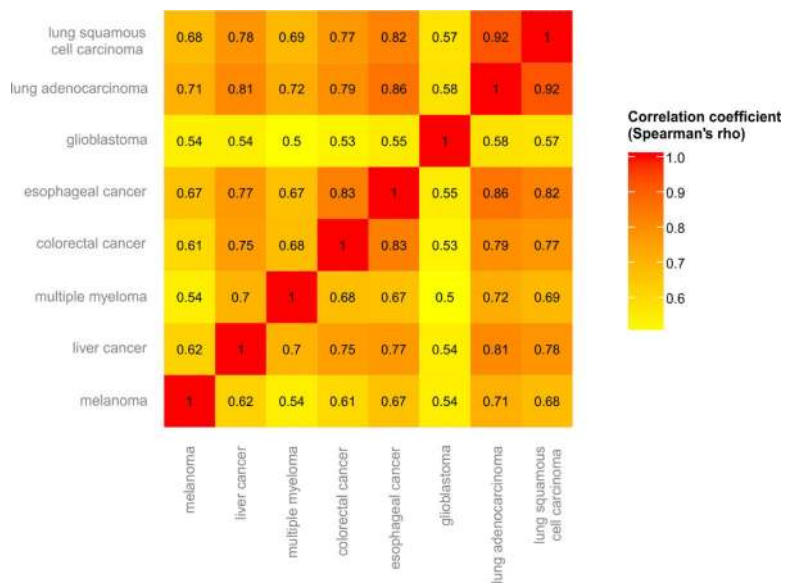
### **Prediction of tissue of origin for individual cancer genomes**

For each individual cancer genome we predicted the density of mutations using Random Forest regression with 10-fold cross-validation. We used the full set of features and determined the top 20 features according to the variable importance measure. We then calculated the enrichment of each tissue type among the top 20 features using the hypergeometric test and chose the tissue showing the most significant enrichment as the most likely tissue of origin for the individual cancer genome. We then calculated the percentage of individual cancers where the assigned tissue of origin matched the predicted tissue of origin.

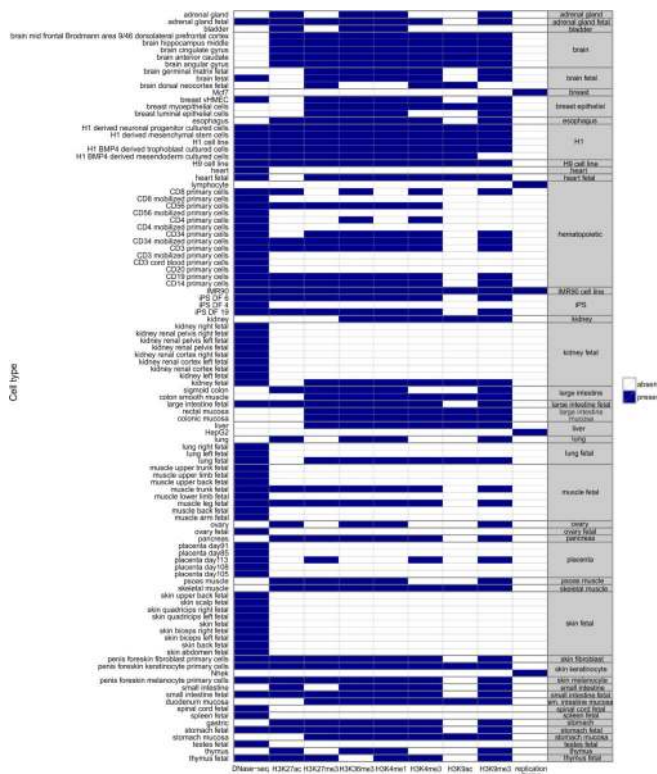
### **Prediction of cell of origin using gene expression**

For each individual cancer we downloaded gene expression data from The Cancer Genome Atlas<sup>15</sup> and calculated the expression of the same genes in the 38 cell types for which mRNA-seq data was available from the Epigenomics Roadmap (Extended Data Table 1). For each cancer we trained a Random Forest regression model in which the gene expression values in cancer were used as the response variable and the gene expression in normal cells as the predictors. We identified the predictor variable, which showed the highest value of variable importance in the model and assigned the corresponding cell type as cell of origin of the cancer.

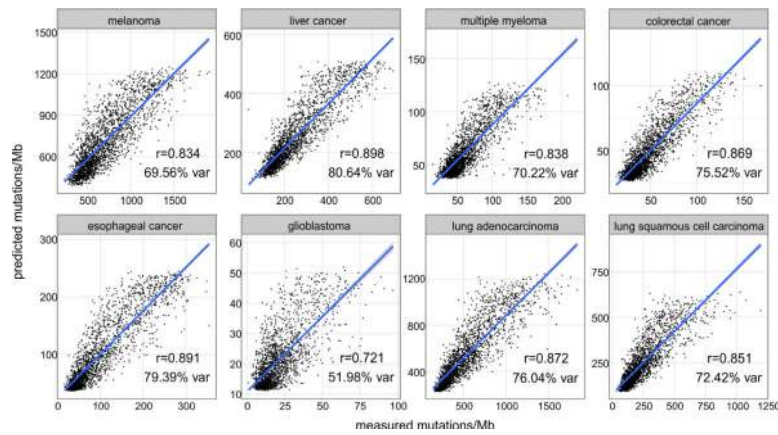
### Extended Data



**Extended Data Figure 1.**  
Correlation of mutation density measured in different cancer types.

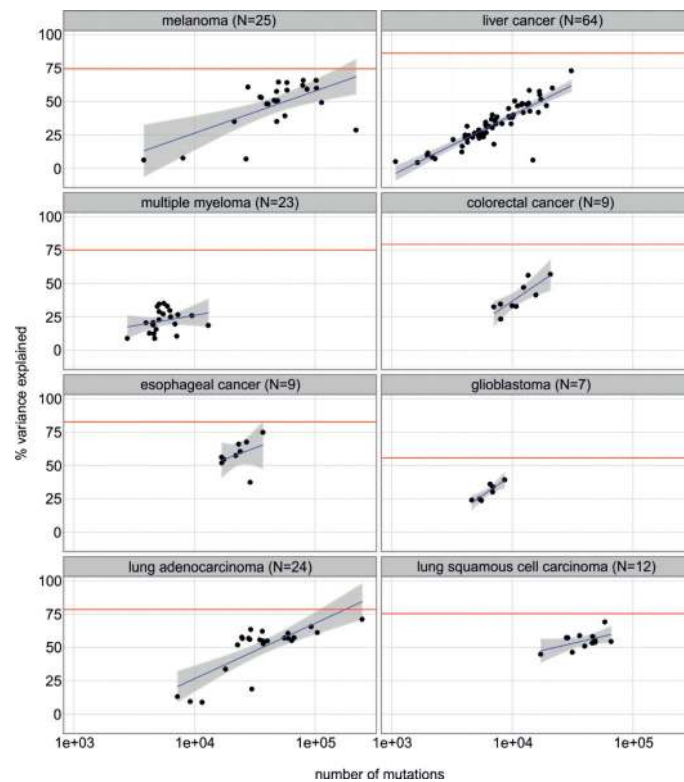


**Extended Data Figure 2.**  
Chromatin features and replication data used in the models.



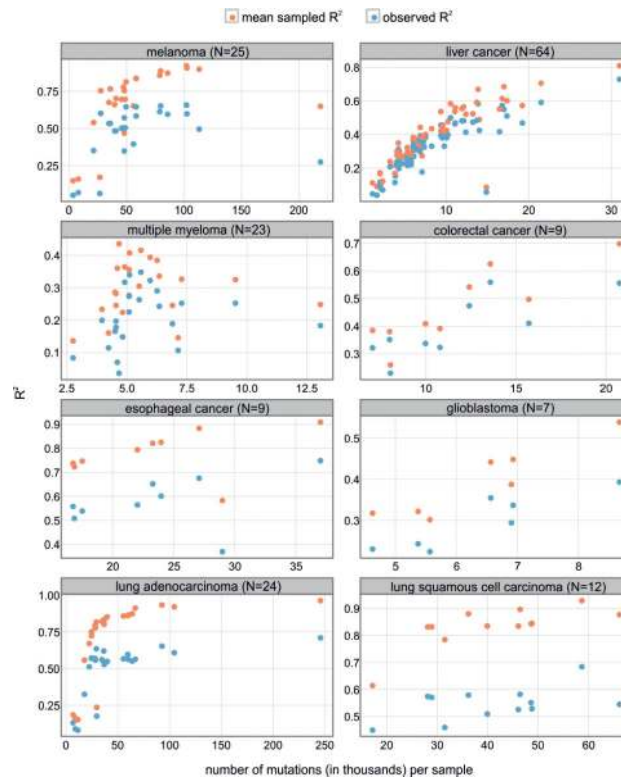
### Extended Data Figure 3.

Scatter plots of the measured number of somatic mutations per Mb in different cancer genomes versus the number of mutations predicted by the Random Forest algorithm. The training set consisted of 10% of the data, while 90% was used to test the predictions.



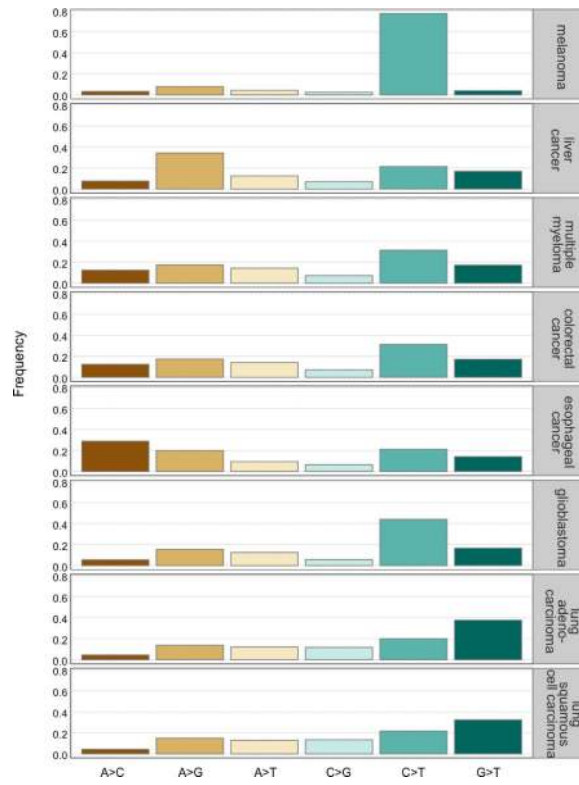
### Extended Data Figure 4.

Prediction accuracy of the models trained on individual cancers as a function of the number of mutations. The red line represents the prediction accuracy of the model used to predict the mutation density of samples pooled by cancer type (sum of all mutations in individual cancers of a certain cancer type). N – number of individual cancers per cancer type.



**Extended Data Figure 5.**

Sampling variance. Red: The squared correlation coefficient ( $R^2$ ) between the observed mutational profile and the profiles predicted by Random Forest. Blue: the maximal attainable variance explained, calculated as the average correlation coefficient squared ( $R^2$ ) between the mutational profiles predicted by Random Forest and 100 simulated mutational profiles modeled as a Poisson distribution with the mean predicted by epigenomic features. N – number of individual cancers per cancer type.



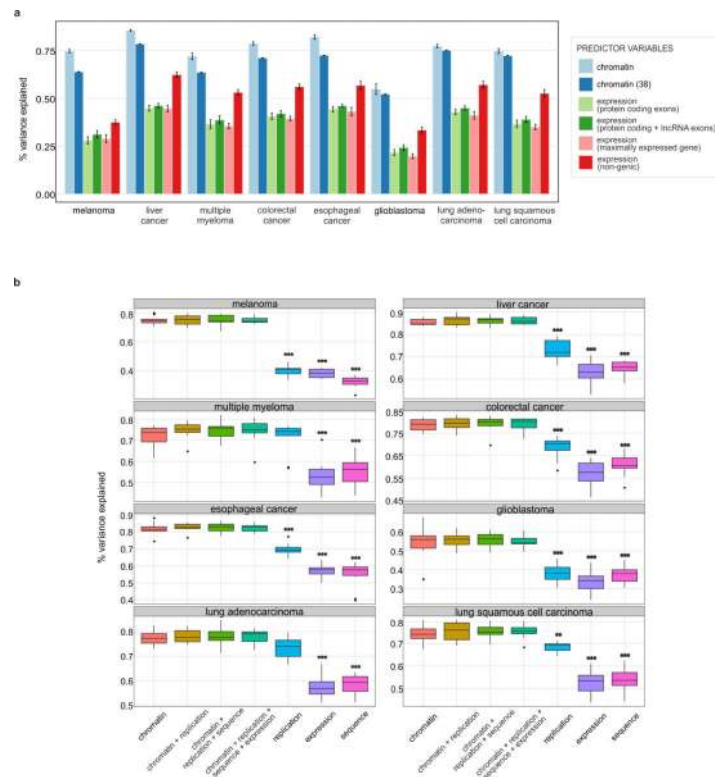
**Extended Data Figure 6.**  
 Frequency of different types of mutations in different cancer types.

Author Manuscript

Author Manuscript

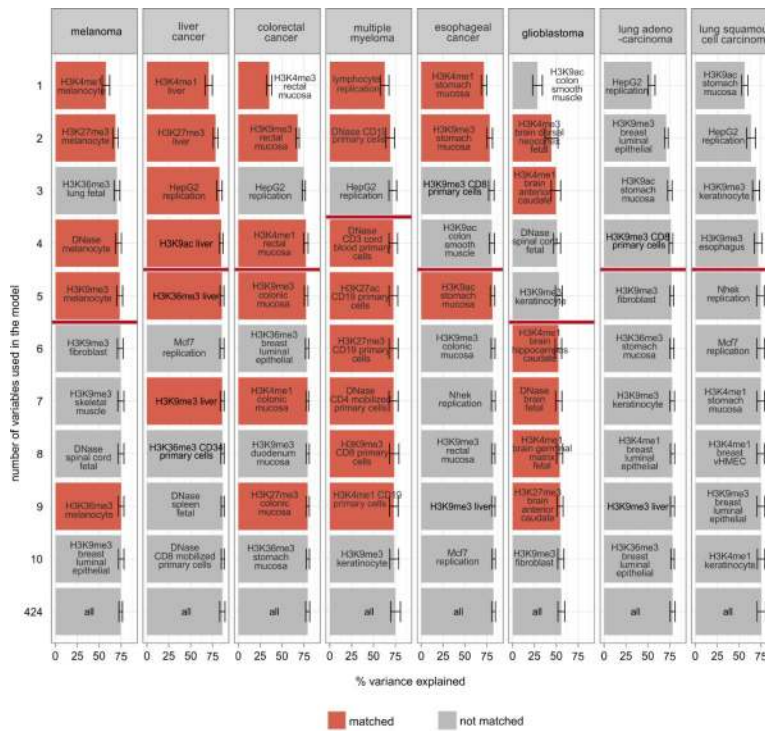
Author Manuscript

Author Manuscript



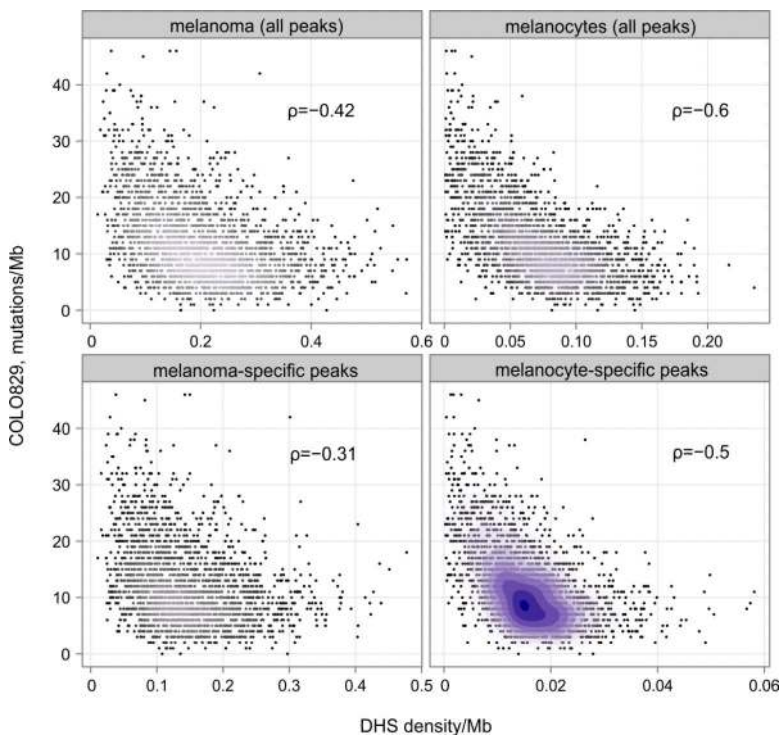
### Extended Data Figure 7.

Prediction accuracy of models obtained using different subsets of predictor variables. (A) Comparison of the prediction accuracy obtained using the full set of chromatin features, 38 chromatin features measured in cell types for which expression data was available, and expression data. Expression in 1MB windows was calculated using mRNA-seq reads mapping to either protein coding exons, protein coding and lncRNA exons, maximally expressed gene or non-genic regions, and normalized by the cumulative length of each of these regions, respectively. Bars represent the mean prediction accuracy; error bars represent standard errors of the mean prediction accuracy estimated using 10-fold cross-validation. (B) Distribution of the percent of variance explained in 10 folds of cross-validation ( $n=10$ ) for models trained on chromatin, replication, expression (non-genic mRNA-seq) or sequence features, or a combination of these subsets of features. Models trained on chromatin features were compared to all other models for a certain cancer type (Wilcoxon rank-sum test). Significant differences, Benjamini–Hochberg-corrected:  $**P < 0.01$ ,  $***P < 0.001$ . Box plots, band inside the box, median; box, first and third quartiles; whiskers, most extreme values within  $1.5 \times$  inter-quartile range from the box; points, outliers.



**Extended Data Figure 8.**

Feature selection by using the backward elimination procedure. For each cancer type, variables are ordered from top to bottom by decreasing importance. Each bar represents the fraction of variance explained by the model using the corresponding bar and all bars above it. The red line indicates the cutoff needed to achieve the prediction accuracy of the full model – 1 s.e.m. For each cancer type, features measured in related cell lines are shown in red.



**Extended Data Figure 9.**  
 The number of mutations per megabase in COLO829 cell line versus DHS density in melanoma cell lines (DHS measured in 11 melanoma cell lines), melanocytes, DHSs specific to melanomas (not observed in melanocytes) and DHSs specific to melanocytes (not observed in melanomas). Correlation is calculated using the Spearman's rank correlation coefficient.

**Extended Data Table 1**

Cell types for which mRNA-seq data was downloaded from Epigenomics Roadmap.

- 
- adrenal gland
  - bladder
  - brain fetal
  - brain germinal matrix fetal
  - brain hippocampus middle
  - breast luminal epithelial cells
  - breast myoepithelial cells
  - breast vHMEC
  - CD34 mobilized primary cells
  - esophagus
  - gastric
  - H1 BMP4 derived mesendoderm cultured cells
  - H1 BMP4 derived trophoblast cultured cells
  - H1 cell line



H1 derived mesenchymal stem cells  
H1 derived neuronal progenitor cultured cells  
IMR90 cell line  
iPS DF 19.11 cell line  
iPS DF 6.9 cell line  
liver  
lung  
lung left fetal  
lung right fetal  
muscle arm fetal  
muscle back fetal  
muscle leg fetal  
muscle trunk fetal  
ovary  
ovary fetal  
pancreas  
penis foreskin fibroblast primary cells  
penis foreskin keratinocyte primary cells  
penis foreskin melanocyte primary cells  
psoas muscle  
sigmoid colon  
small intestine  
spinal cord fetal  
thymus

---

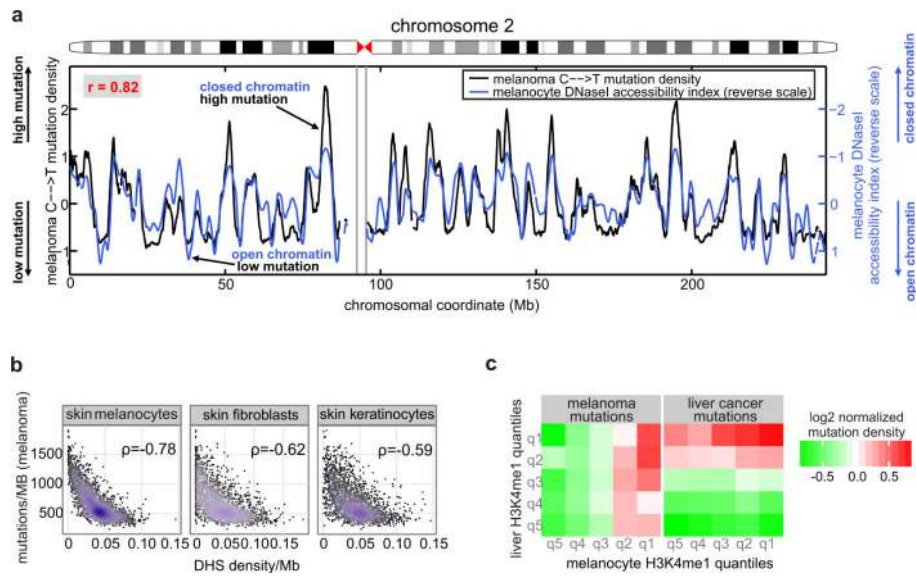
## Acknowledgements

This work was supported by NIH grants R01 MH101244, U54 CA143874 to S.R.S. and U01 ES017156, P01 HL53750, U54 HG007010 to J.A.S. R.K and K.V. acknowledge the Integra-Life Seventh Framework Program (grant number 315997) and the EMBO Young Investigator Program (Installation grant 1431/2006 to KV).

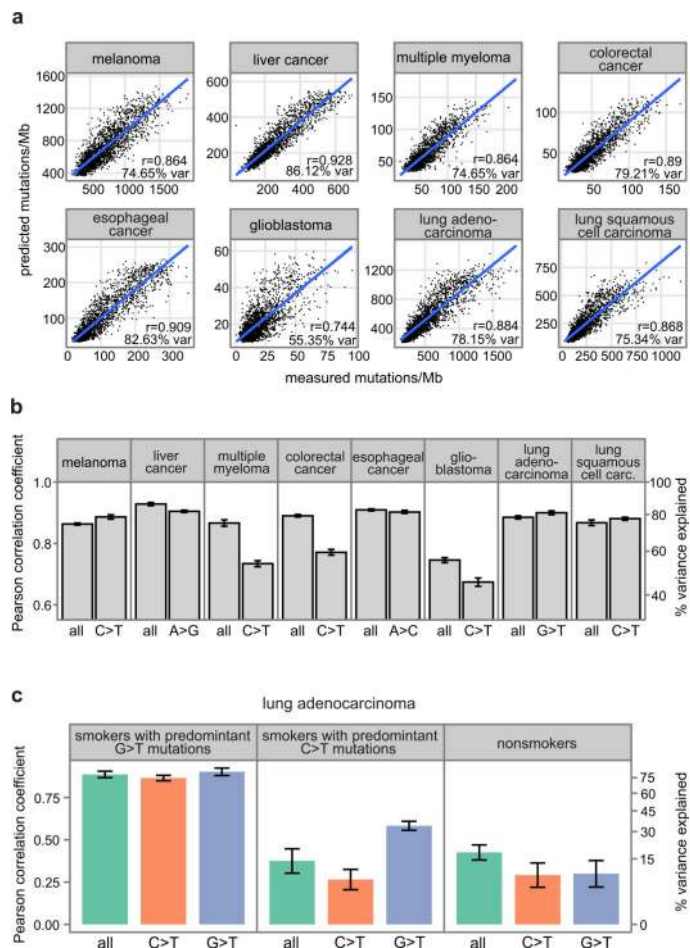
## References

1. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 2013; 499:214–218. doi:10.1038/nature12213. [PubMed: 23770567]
2. Hodgkinson A, Chen Y, Eyre-Walker A. The large-scale distribution of somatic mutations in cancer genomes. *Human mutation*. 2012; 33:136–143. doi:10.1002/humu.21616. [PubMed: 21953857]
3. Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nature communications*. 2013; 4:1502. doi:10.1038/ncomms2502.
4. Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*. 2012; 488:504–507. doi:10.1038/nature11273. [PubMed: 22820252]
5. Woo YH, Li WH. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nature communications*. 2012; 3:1004. doi:10.1038/ncomms1982.
6. Zhu J, et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell*. 2013; 152:642–654. doi:10.1016/j.cell.2012.12.033. [PubMed: 23333102]

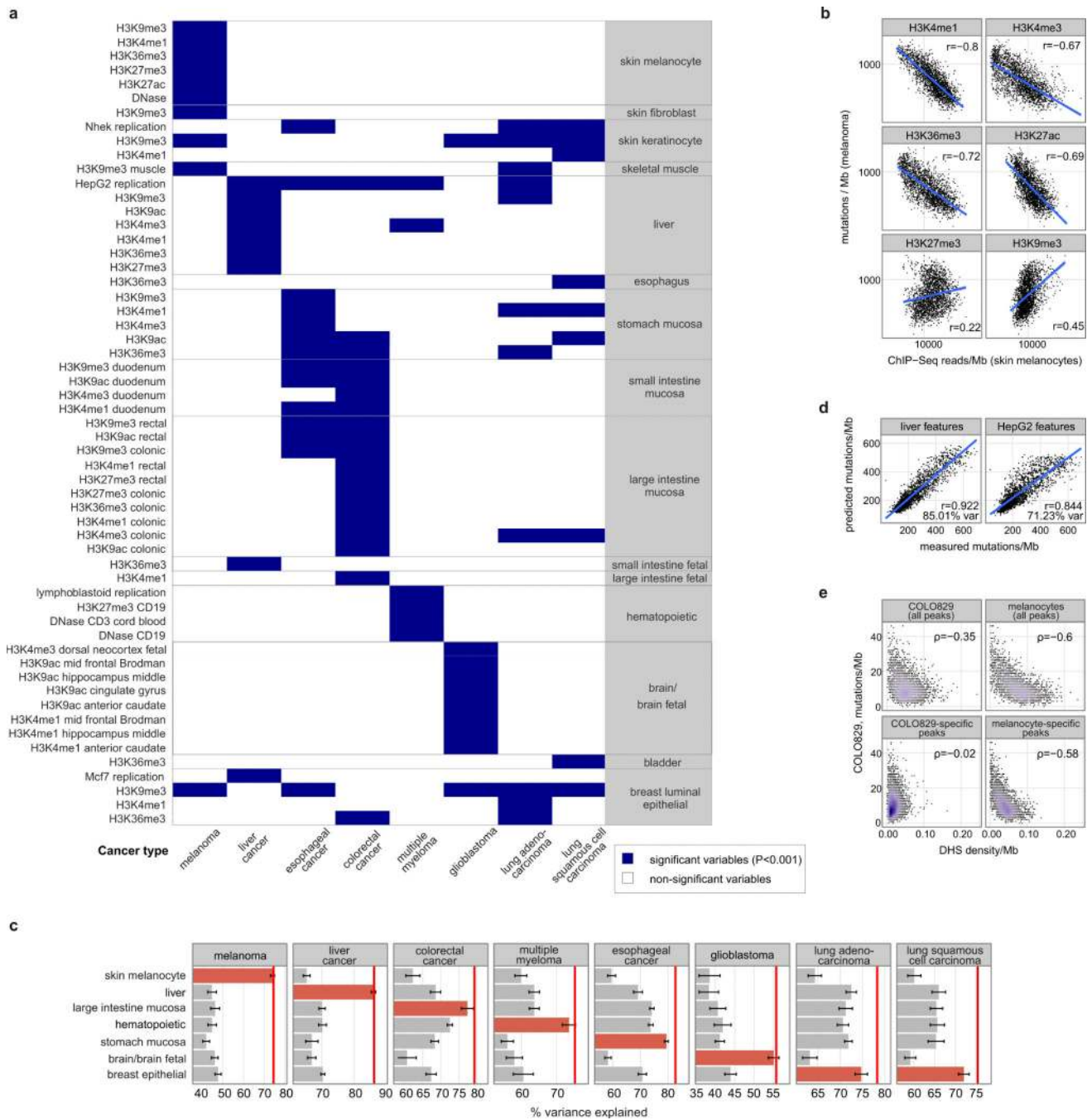
7. Thurman RE, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012; 489:75–82. doi:10.1038/nature11232. [PubMed: 22955617]
8. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. doi:10.1038/nature11247. [PubMed: 22955616]
9. Ryba T, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*. 2010; 20:761–770. doi:10.1101/gr.099655.109. [PubMed: 20430782]
10. Berger MF, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature*. 2012; 485:502–506. doi:10.1038/nature11071. [PubMed: 22622578]
11. Chapman MA, et al. Initial genome sequencing and analysis of multiple myeloma. *Nature*. 2011; 471:467–472. doi:10.1038/nature09837. [PubMed: 21430775]
12. Imielinski M, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*. 2012; 150:1107–1120. doi:10.1016/j.cell.2012.08.029. [PubMed: 22980975]
13. Totoki Y, et al. High-resolution characterization of a hepatocellular carcinoma genome. *Nature genetics*. 2011; 43:464–469. doi:10.1038/ng.804. [PubMed: 21499249]
14. Bass AJ, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nature genetics*. 2011; 43:964–968. doi:10.1038/ng.936. [PubMed: 21892161]
15. Brennan CW, et al. The somatic genomic landscape of glioblastoma. *Cell*. 2013; 155:462–477. doi:10.1016/j.cell.2013.09.034. [PubMed: 24120142]
16. Dulak AM, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature genetics*. 2013; 45:478–486. doi:10.1038/ng.2591. [PubMed: 23525077]
17. TCGA. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012; 489:519–525. doi:10.1038/nature11404. [PubMed: 22960745]
18. Hansen RS, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America*. 2010; 107:139–144. doi:10.1073/pnas.0912402107. [PubMed: 19966280]
19. Pleasance ED, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*. 2010; 463:191–196. doi:10.1038/nature08658. [PubMed: 20016485]
20. Alexandrov LB, et al. Signatures of mutational processes in human cancer. *Nature*. 2013 doi:10.1038/nature12477.
21. Pleasance ED, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature*. 2010; 463:184–190. doi:10.1038/nature08629. [PubMed: 20016488]
22. Reid BJ, Li X, Galipeau PC, Vaughan TL. Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis. *Nature reviews. Cancer*. 2010; 10:87–101. doi:10.1038/nrc2773.
23. Hudson TJ, et al. International network of cancer genome projects. *Nature*. 2010; 464:993–998. doi:10.1038/nature08987. [PubMed: 20393554]
24. Cibulskis K, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*. 2013; 31:213–219. doi:10.1038/nbt.2514.
25. Neph S, et al. BEDOPS: high-performance genomic feature operations. *Bioinformatics*. 2012; 28:1919–1920. doi:10.1093/bioinformatics/bts277. [PubMed: 22576172]
26. Koren A, et al. Differential Relationship of DNA Replication Timing to Different Forms of Human Mutation and Variation. *American journal of human genetics*. 2012 doi:10.1016/j.ajhg.2012.10.018.
27. Breiman L. Random Forests. *Machine learning*. 2001; 45:5–32. doi:10.1023/A:1010933404324.
28. Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics*. 2010; 26:1340–1347. doi:10.1093/bioinformatics/btq134. [PubMed: 20385727]



**Figure 1.** Mutation density in melanoma is associated with individual chromatin features specific to melanocytes. (a) The density of C>T mutations in melanoma alongside a 100kb window profile of melanocyte chromatin accessibility (“DNase I accessibility index”; shown in normalized, reverse scale; high values correspond to less accessible chromatin and vice versa). (b) The number of mutations per megabase in melanoma versus DHS density, for three types of skin cells. (c) The normalized density of mutations in liver cancer and melanoma genomes as a function of density quintiles of H3K4me1 marks in liver cells and in melanocytes. For both cancer genomes, mutation density depends only on H3K4me1 marks measured in the cell of origin.

**Figure 2.**

Predicting local mutation density in cancer genomes using Random Forest regression trained on 424 epigenomic profiles. Pearson correlation between observed and predicted mutation densities along chromosomes is shown. (a) Actual versus predicted mutation densities in eight cancers. (b, c) Prediction accuracy represented as mean  $\pm$  s.e.m (estimated using 10-fold cross-validation). Panels show prediction accuracy for all mutations and for nucleotide changes predominant in the corresponding cancer (b), and prediction accuracy in lung adenocarcinoma genomes stratified by smoking history and predominant nucleotide changes (G>T or C>T) (c).



**Figure 3.** Epigenomic features that significantly contribute to the prediction of local mutation density. (a) Features (blue rectangles) significantly contributed to the predictions in at least one cancer type (see Methods). (b) Melanoma mutation density versus the density of chromatin modifications in melanocytes. (c) Prediction accuracy (mean ± s.e.m estimated using 10-fold cross-validation) of models separately trained on features from different tissues for each cancer type. Red bars: tissues with the highest prediction accuracy. Red line: prediction accuracy when using all 424 epigenetic features. (d) Comparison of predictions accuracies

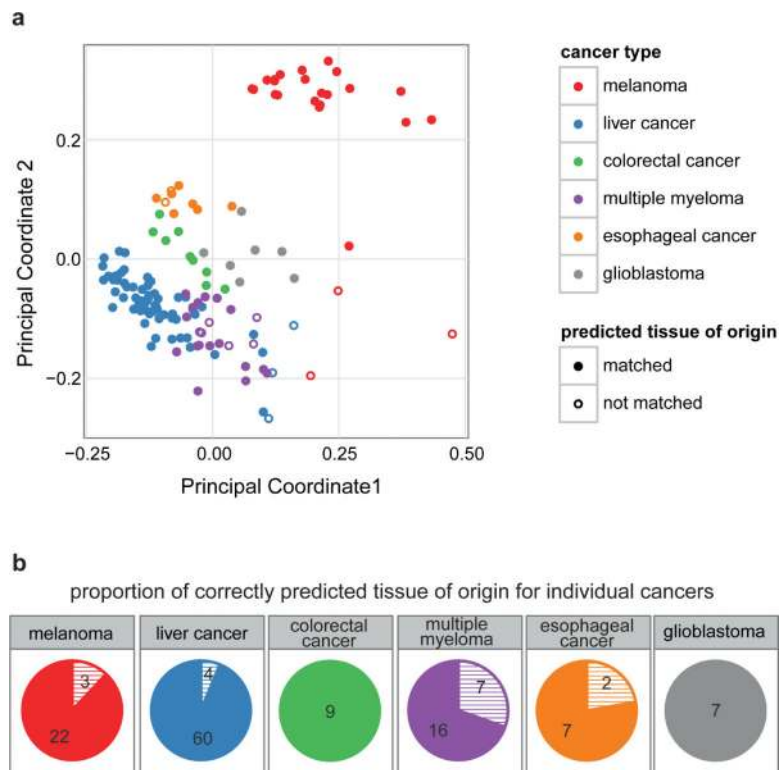
of liver cancer mutation density from features of normal liver cells vs. cancer cells (HepG2). (E) Mutation density in COLO829 melanoma cell line versus DHS density in COLO829, melanocytes, DHSs specific to COLO829 (not observed in melanocytes) and DHSs specific to melanocytes (not observed in COLO829). Spearman's rank correlation coefficient is given for each comparison.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



**Figure 4.**

Analysis of individual cancer genomes and prediction of cell type of origin. (a) Principal coordinate analysis (PCOA) of the distribution of mutations in individual cancer genomes. Filled circles represent cancers for which the correct cell type of origin was identified. (b) The accuracy of cell type of origin prediction for individual cancer genomes: the number of cancer samples that were assigned to the correct (solid colors) or incorrect (textures) cell types of origin based on their mutation profile.