Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer

Authors: Young Seok Ju[1], Ludmil B Alexandrov[1], Moritz Gerstung[1], Inigo Martincorena[1], Serena Nik-Zainal[1], Manasa Ramakrishna[1], Helen R Davies[1], Elli Papaemmanuil[1], Gunes Gundem[1], Adam Shlien[1], Niccolo Bolli[1], Sam Behjati[1], Patrick S Tarpey[1], Jyoti Nangalia[1,2,3], Charles E Massie[1,2,3], Adam P Butler[1], Jon W Teague[1], George S Vassiliou[1,2,3], Anthony R Green[2,3], Ming-Qing Du[2], Ashwin Unnikrishnan[4], John E Pimanda[4], Bin Tean Teh[5,6], Nikhil Munshi[7], Mel Greaves[8], Paresh Vyas[9], Adel K El-Naggar[10], Tom Santarius[2], V Peter Collins[2], Richard Grundy[11], Jack A Taylor[12], D Neil Hayes[13], David Malkin[14], ICGC Breast Cancer Group[1]†, ICGC Chronic Myeloid Disorders Group[1]†, ICGC Prostate Cancer Group[1,8,15]†, Christopher S Foster[16], Anne Y Warren[2], Hayley C. Whitaker[15], Daniel Brewer[8,17], Rosalind Eeles[8], Colin Cooper[8,17], David Neal[15], Tapio Visakorpi[18], William B Isaacs[19], G Steven Bova[18], Adrienne M Flanagan[20,21], P Andrew Futreal[1,10], Andy G Lynch[15], Patrick F Chinnery[22], Ultan McDermott[1,2], Michael R Stratton[1] and Peter J Campbell[1,2,3]*

Affiliations:

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK

[2]Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

[3]Department of Haematology, University of Cambridge, Cambridge, UK

[4]Lowy Cancer Research Centre, University of New South Wales, Sydney, Australia

[5]Laboratory of Cancer Epigenome, National Cancer Centre, Singapore

[6]Duke-NUS Graduate Medical School, Singapore

[7]Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[8]Institute of Cancer Research, Sutton, London, UK

[9]Weatherall Institute for Molecular Medicine, University of Oxford, Oxford UK

[10]MD Anderson Cancer Center, Houston, Texas, USA

[11]Children's Brain Tumour Research Centre, University of Nottingham, Nottingham, UK

[12]National Institute of Environmental Health Sciences, NIH, North Carolina USA

[13]University of North Carolina, North Carolina, USA

[14]Hospital for Sick Children, University of Toronto, Toronto, Canada

[15]Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK

[16]University of Liverpool and HCA Pathology Laboratories, London, UK

[17]University of East Anglia, Norwich, UK

[18]Institute of Biomedical Technology and BioMediTech, University of Tampere, Tampere, Finland

[19]Johns Hopkins University, Baltimore, Maryland, USA

[20]Royal National Orthopaedic Hospital, Middlesex, UK

[21]UCL Cancer Institute, University College London, London, UK

[22]Wellcome Centre for Mitochondrial Research, Institute of Genetic Medicine, Newcastle University, Newcastle-upon-Tyne, UK

† Participants listed in appendix

Contact info:

Correspondence to Dr. Peter J Campbell
Cancer Genome Project,
Wellcome Trust Sanger Institute,
Hinxton CB10 1SA, United Kingdom.
Phone: +44 1223 494745
e-mail: pc8@sanger.ac.uk

ABSTRACT

Recent sequencing studies have extensively explored the somatic alterations present in the nuclear genomes of cancers. Although mitochondria control energy metabolism and apoptosis, the origins and impact of cancer-associated mutations in mitochondrial DNA (mtDNA) are unclear. Here, we analysed somatic alterations in mtDNA from 1,675 tumors across 31 histologies. We identified 1,907 somatic substitutions, which exhibited dramatic replicative strand bias, predominantly C>T and A>G on the mitochondrial heavy strand. This strand-asymmetric signature differs from those found in nuclear cancer genomes but matches the inferred germline process shaping primate mtDNA sequence content. Numbers of mtDNA mutations showed considerable heterogeneity across tumor types. Missense mutations were selectively neutral and often gradually drifted towards homoplasmy over time. In contrast, mutations resulting in protein truncation undergo negative selection and were almost exclusively heteroplasmic. Our findings indicate that the endogenous mutational mechanism has far greater impact than any other external mutagens in mitochondria, and is fundamentally linked to mtDNA replication.

INTRODUCTION

All cancers result from somatic mutations in their genomes. Beyond the ~3,200Mb of nuclear genomic DNA, human cells have hundreds to thousands of mitochondria present in every cell, each carrying one or a few copies of the 16,569bp circular mitochondrial genomes (Koppenol et al., 2011; Legros et al., 2004; Smeitink et al., 2001). In addition to their role in cellular energy balance through oxidative phosphorylation, mitochondria are involved in many essential cellular functions including modulation of oxidation-reduction status, contribution to cytosolic biosynthetic precursors and initiation of apoptosis. Mitochondria in eukaryotic cells evolved by endosymbiosis from a free-living α-proteobacterium (Gray et al., 1999). Over two billion years of co-evolution, many ancestral mitochondrial genes have transferred to the nucleus (Calvo and Mootha, 2010; Falkenberg et al., 2007; Wallace, 2012). What remains in the mitochondrial genome is distinctive for the striking asymmetry between the two complementary mtDNA strands in terms of nucleotide content and gene distribution (Andrews et al., 1999). The heavy (H) strand is guanine-rich (C/G=0.4) and is the template from which most mitochondrial proteins (12 out of 13) are transcribed, whereas only one protein-coding gene, *MT-ND6*, is transcribed from the correspondingly cytosine-rich light (L) strand.

Mutations in the mitochondrial genome cause inherited disease (Chinnery, 1993), with a maternal inheritance pattern because only eggs contribute mitochondria to the zygote. The penetrance of inherited mitochondrial disease is determined stochastically by both the random assortment of mutated versus wild-type mitochondrial genomes during meiosis and random drift during the early cell divisions after fertilization. In cancer, the role of somatically acquired mtDNA mutations is controversial. Although cancer-specific mutations have been previously

reported (Brandon et al., 2006; Chatterjee et al., 2006; He et al., 2010; Larman et al., 2012; Polyak et al., 1998), the limited sample size or poor sensitivity of capillary sequencing for heteroplasmic mutations have not allowed a comprehensive analysis of the mutational signatures of mitochondrial mutations nor their likely functional significance. It has long been proposed that mitochondria might contribute to cancer development given their fundamental importance to cellular biology (Wallace, 2012). Previous reports suggested mitochondrial somatic mutations might be under positive selection and thus contribute to cancer development, but the small numbers of reported mutations render this conclusion uncertain (Brandon et al., 2006; Chatterjee et al., 2006; Larman et al., 2012; Schon et al., 2012). Nonetheless, the hypothesis of functionally relevant mitochondrial mutations is an appealing one because cancer cells have greatly increased energy demands over normal cells, and demonstrate a switch from aerobic glycolysis in mitochondria to lactic acid fermentation in the cytosol (the Warburg effect) (Hanahan and Weinberg, 2011; Koppenol et al., 2011).

In each cell cycle, the replicating genome is at risk of *de novo* mutations, which can promote the development of cancer. These mutations may be generated by intrinsic cellular errors during DNA replication or repair or through exposure to mutagens, such as reactive oxygen species, tobacco smoke and ultraviolet light (Pleasance et al., 2010a; Pleasance et al., 2010b). Recently, >20 mutational signatures operative in cancers have been identified in the nuclear genome (Alexandrov et al., 2013). Whether any of these mutational processes also affect the mitochondrial genome has not been studied. Furthermore, whether there are mtDNA-specific mutational processes in somatic cells remains unclear, although the many unique features of mtDNA replication and repair, coupled with the high concentration of reactive oxygen species

generated by the electron transport chain, could be associated with distinctive mutation signatures.

In this study, we compare 1,675 cancer and paired-normal mtDNA sequences across 31 tumor types using massively parallel DNA sequencing technologies to obtain a systematic and unbiased catalogue of somatic mitochondrial mutations. We find that mtDNA mutations are almost exclusively the product of a mutational process that is specific to mitochondria and probably linked to the unique mechanism of genome replication these organelles employ. We find no evidence for positive selection of mitochondrial mutations during oncogenesis, suggesting that they confer no clonal advantage on the nascent cancer cells.

RESULTS

mtDNA Sequencing and Mutation Calling

We extracted the mtDNA sequences from 704 whole-genome and 971 whole-exome sequencing data generated on primary cancers and compared them with mtDNA sequences from their matched normal samples. Given the abundance of mtDNA per cancer cell, a standard coverage of 30-40x in the nuclear genome provides significantly greater coverage of the mitochondrial genome (average read-depth = 7901.0x), enabling accurate identification of somatic mutations including rare heteroplasmic variants. We also assessed whether whole exome sequencing could be used to identify mtDNA mutations from off-target reads derived from the mitochondrial genome. We found an average read-depth of 92.1x across the mitochondrial genome in exome studies. From 139 samples in which we had both exome and whole genome sequencing data, the overall read depths correlated strongly (R^2=0.59, Figure 1-figure supplement 1) as did variant allele fractions for mtDNA somatic mutations (R^2=0.97, Figure 1-figure supplement 2). Validation experiments suggested the sensitivity of whole-exome sequencing for detection of mtDNA somatic mutations to be 71.4% compared to whole-genome sequencing (Figure 1-figure supplement 3 and Materials and Methods, "Off-target mtDNA reads in whole-exome sequencing" and "DNA cross-contamination").

To reduce potential false-positive calls of mtDNA somatic mutations, we only report variants called with an allele fraction of >3%. This eliminates the risk of miscalls due to mtDNA-derived pseudogenes in the nucleus (NuMTs) because mtDNA copy numbers are 100-1000 times higher than nuclear genomes in human somatic cells and the sequence homology between mtDNA and NuMTs presented in the human reference genome is generally <95% (in 96 out of 101 NuMTs

with length greater than 300bp). Furthermore, pairwise comparison between cancer and matched normal mtDNAs from the same individual further minimizes the contamination of NuMTs in the mutation calling.

The Catalog of mtDNA Somatic Mutations

In total, 1,675 tumor-normal pairs across 31 tumor types were analysed (Table 1 and Supplementary file 1). For 61 of these patients, we had sequencing data available from multiple sites of the primary cancer, several time-points or matched primary cancers and metastases (a total of 73 such cancer samples), allowing us to study the timing of mtDNA mutations in cancer evolution (Supplementary file 1). We identified 1,907 somatic mtDNA substitutions (Figure 1 and Supplementary file 2). In contrast to inherited polymorphisms (n=38,706, available at Supplementary file 2), which were almost always homoplasmic in both the cancer and counterpart normal, the variant allele fractions (VAFs) of these somatic substitutions were highly variable in the cancer, ranging from our detection threshold (3%) to homoplasmy (100%). Of these 1,907 somatic substitutions, 1,209 (63.4%) were not registered in the databases of mtDNA common polymorphism (Ingman and Gyllensten, 2006; Levin et al., 2013). In comparison, when we examined substitutions found in both the tumor and the normal samples from a patient, only 21 (0.05%) were not registered in the polymorphism databases, a significantly different fraction from the tumor-only variants ($p<10^{-10}$; chi-squared test). We found 595 (31.2%) recurrent mutations that can be collapsed onto 246 mtDNA positions, which is a 6.9-fold higher level of recurrence than expected by chance ($p<10^{-10}$). This suggests that the generation or fixation of mtDNA mutations is not random, but influenced by factors such as the underlying mutational process or positive selection.

Of the 1,675 cancer samples, 976 (58.3%) harbored at least one somatic substitution and 521 (31.1%) had multiple substitutions, ranging from 2 to 7 (Figure 2A). In those with multiple substitutions, 72 pairs of mutations were sufficiently close to phase (Nik-Zainal et al., 2012b) such that we could determine whether they were linked on the same mtDNA genome or were on different copies. We found that 45 (62.5%) pairs of mutations were linked on the same mtDNA genome (Supplementary file 3 and Figure 2-figure supplement 1). Furthermore, of these linked mutations, 33 showed a clear temporal order: that is, one mutation was demonstrably subclonal to the other. This is rather unexpected, since each somatic cell has 100-1,000 copies of the mitochondrial genome, and we might anticipate that random mutations would, on average, affect different copies. That many pairs of mutations are phased on the same mtDNA genome and yet show a clear subclonal relationship suggests that they occur sufficiently separated in time to allow the mitochondrial genome carrying the earlier mutation to drift towards a substantial fraction of all genomes in that cell before the second mutation occurs, consistent with a previous report (De Alwis et al., 2009).

The number of somatic mtDNA substitutions varied significantly according to tumor type ($p=4.4\times10^{-52}$) after correcting for confounding variables such as sequencing coverage: gastric, hepatocellular, prostate and colorectal cancers had the highest numbers of mtDNA substitutions (Figure 2B). In contrast, hematologic cancers (acute lymphoblastic leukaemia, myeloproliferative disease and myelodysplastic syndrome) had fewer mutations. Several possible explanations could underpin these differences across tumor types. It could be that the mutation rates differ across cell lineages; it could be that selection pressures shape the number of

mutations; or the number of mtDNA genome generations could differ across cell lineages. Of these explanations, we believe that the second is unlikely because, as we shall see, positive selection is not a major component of mitochondrial mutations. Interestingly, we find a positive correlation between the number of mtDNA somatic mutations and age at diagnosis in breast cancers ($p=0.0004$; Figure 2C), in keeping with the idea that the number of mitochondrial generations is linked to mutation burden. The mutational burden of an established cancer represents the accumulated variation acquired in the lineage of cell divisions from fertilized egg to transformed cell, and will include events acquired in normal development and homeostasis as well as those acquired during tumorigenesis (Stratton et al., 2009). Interestingly, mtDNA mutations have been found at high rates in normal colonic crypt cells (Ericson et al., 2012; Taylor et al., 2003). Given that we find high burdens of mutations in colonic tumors as well, the differences we see across tumor types may arise from pre- or post-transformation differences in mtDNA burden across tissues.

Extracting mtDNA Mutational Signatures

With respect to signatures of somatic substitutions, C>T and T>C transitions constituted 90.9% of all the 1,907 substitutions (Figure 1) among the six classes of possible base substitutions. To characterize this aggregated signature of mtDNA cancer specific mutations in more detail, we looked for the presence of mtDNA strand bias between the complementary H and L strands of mtDNA. The two main substitution classes showed an extreme level of mtDNA strand bias. 84.1% of the C>T transitions were on the H strand. This level of strand bias occurred despite the fact that cytosine is 2.4-fold less common on the H than the L strand, so the C>T substitution rate is 12.6-fold higher on the H strand. By contrast, 76.8% of the T>C transitions were on the L

strand despite its lower thymine content (1.3-fold less than the H strand). This implies that the T>C mutation rate on the L strand is 4.2-fold higher than on the H strand.

We then examined the sequence context in which these mutations occurred by examining the bases immediately 5' and 3' to the mutated bases. This generates 96 possible mutation classes (the 6 substitution classes multiplied by the 16 combinations of immediate 5' and 3' nucleotides). Both C>T and T>C mutations showed highly distinctive sequence contexts. $C_H>T_H$ substitutions (i.e. C>T mutations on the H strand) were enriched for the Np<u>C</u>pG trinucleotide context (8- to 15-fold more frequent than expected by chance; Figure 3A). By contrast, $T_L>C_L$ substitutions (i.e. T>C mutations on the L strand) showed 5- to 8-fold enrichment in Np<u>T</u>pC. This strand-asymmetric mutational signature is not similar to any of the 21 cancer associated mutational signatures recently identified from the nuclear DNA of 30 different cancer types (Alexandrov et al., 2013).

Of the 18 tumor types that presented at least 25 mtDNA somatic substitutions in this study, the mutational signatures were broadly consistent across tumor types (Figure 3B), with the exception that multiple myeloma had a somewhat higher rate of $T_H>C_H$ changes than other histologies ($p=8.1\times10^{-6}$). Thus, in contrast to the mutational signatures found in nuclear genomes, where there is striking heterogeneity both across tumor types and across individuals within a tumor type (Alexandrov et al., 2013), the mutational profile in the mitochondrial genome of somatic cells is remarkably homogeneous.

Replication-coupled Mutational Process in Mitochondria

The major known cause of mutational strand bias in nuclear DNA is transcription-coupled nucleotide excision repair, where DNA lesions on the transcribed (non-coding) strand are more frequently repaired (Alexandrov et al., 2013). However, we find that the strand bias always favors $C_H>T_H$ and $T_L>C_L$ whether the gene is transcribed from the H strand or from the L strand (Figure 3-figure supplement 1). This is not compatible with transcription-coupled repair, for which the direction of strand bias is fundamentally dictated by which strand is transcribed.

Instead, the mtDNA mutational strand bias reported here appears to be driven by differences in replication between the two strands. mtDNA replication harbors substantial strand asymmetry between the H and L strands: mtDNA replication initiates from an origin of replication (O_H) in the D-Loop, with the nascent H and the L strand replicating as leading and lagging strand, respectively (Clayton, 1982; Falkenberg et al., 2007; Holt and Reyes, 2012). We observed that C>T substitutions were prevalent in the leading (heavy) strand, whereas T>C substitutions were found in the lagging (light) strand (Figure 1). Remarkably, this strand bias was reversed in the D-Loop itself (Figures 1 and 3C), further suggesting that the mtDNA somatic mutations are replication-coupled: according to a recently proposed bidirectional model of mtDNA replication (Holt and Reyes, 2012; Yasukawa et al., 2006; Yasukawa et al., 2005), mtDNA replication is also able to initiate from the so-called Ori-b site, typically located around genomic position 16,197 and proceeds on both strands away from the origin (Figure 1). Replication of the nascent H strand continues unimpeded like the traditional model, but the nascent L strand terminates at the so-called O_H site, typically around mtDNA position 191bp. Under this model, then, the

leading and lagging strand are reversed in the few hundred base-pairs of the D-Loop, which is consistent with the reversed mutational signature in this region (Figures 1 and 3C).

Equivalent Mutational Signature during Human mtDNA Evolution

It is not entirely straightforward to infer the mutational signatures operating on the mitochondrial genome in the germline. *De novo* mutations are generally rare and often discovered because they cause disease; distinguishing the ancestral base and the derived base is challenging for single nucleotide polymorphisms; and comparative mtDNA genomics across species extends over considerable evolutionary time. In contrast, because ancestral and derived states are defined for tumor-normal pairs, a much clearer picture emerges of the somatic mtDNA mutation signature. We therefore assessed whether the signature that emerges for somatic mitochondrial mutations could extend to explain sequence composition of the human mtDNA genome.

It appears that the mutational mechanism which has generated the $C_H>T_H$ and $T_L>C_L$ signature in cancer mtDNA is equivalent to the one that has been operating during evolution of human germline mtDNA (Nikolaou and Almirantis, 2006). This manifests as the depletion of certain codons in the reference human mtDNA sequence through the action of the $C_H>T_H$ and $T_L>C_L$ mutational process over time (Figure 4A). For example, the GC<u>G</u> triplet codon (Alanine) appears to have been replaced by its synonymous GC<u>A</u> codon (due to $C_H>T_H$ ($G_L>A_L$)), with the former being 15.8-fold less frequently observed in the 12 mtDNA protein-coding genes that are transcribed from the H strand (and encoded on the L strand). All 32 synonymous codon pairs present the same tendency. Consistent with this interpretation, the gene transcribed from the L

strand (*MT-ND6*) demonstrates the opposite direction of skew. Further analyses of mtDNA codon usage from seven animal species suggest that the $C_H>T_H$ and $T_L>C_L$ mutational pressure may be characteristic of vertebrates, and primates in particular (Figure 4-figure supplement 1).

To quantify whether the somatic mutational signature we have defined can fully explain the trinucleotide frequency of human mtDNA, we performed evolutionary simulations. First, we simulated the evolution of a random DNA sequence under the mutational signature described here. By mutational pressure alone, the random sequence starts losing certain hyper-mutable trinucleotides until eventually reaching a stationary sequence composition. The actual sequence composition of the human mitochondrial genome strongly resembles this stationary distribution (Pearson's $r=0.83$; $p<0.0001$; Figure 4-figure supplement 2). In a second simulation, a random sequence encoding the exact amino acid sequence of the reference mitochondrial genome was evolved by synonymous mutations under the observed mtDNA signature until reaching a stationary sequence composition (mutation-selection equilibrium). These simulations also eventually approximate the observed human mitochondrial genome (Pearson's $r=0.96$, $p<0.0001$; Figure 4B). These analyses strongly suggest that the mitochondrial mutation signature observed in cancer cells closely reflects the mutation signature active in the germline, which has continuously shaped the mitochondrial genome during human evolution.

Negative Selection on Truncating mtDNA Mutations and tRNA Anticodons

Next, we assessed the functional impact of somatic mtDNA mutations. Of the 1,907 substitutions, 1,153 (60.5%) were in the 13 protein-coding genes. These include 63 nonsense, 4

stop-lost, 878 missense and 208 silent substitutions (Supplementary file 2). In addition, out of 251 indels we observed, 110 occurred within protein-coding genes (Supplementary file 2). Of the missense substitutions, 245 (27.9%) were recurrent, affecting 107 distinct mtDNA sites. Although this very high level of mutation clustering could, at first sight, be interpreted as evidence for positive selection, we found that silent substitutions were also frequently recurrent (28 recurrent variants in 13 mtDNA sites), with no substantial difference in recurrence rates between silent and missense mutations (p=0.19; Figure 5-figure supplement 1). We believe this recurrence to be the consequence of a high mtDNA mutation rate with restricted mutational signature (C_H>T_H and T_L>C_L). Independently recurring mutations in human germline mtDNA are well described across human evolution (Levin et al., 2013).

The ratio of somatic missense to silent substitutions (Rms:s) is apparently higher (4.2, 878/208) than that observed for cancer-associated somatic mutations in nuclear DNA (generally around 2:1 to 3:1 across tumor types (Greenman et al., 2007; Nik-Zainal et al., 2012a). At face value, this again could be interpreted as evidence for positive selection. However, as described above, the somatic mtDNA mutational signature shows extreme strand asymmetry and the same mutational signature has been operative in the germline over evolutionary time. Thus, the dominant mutational signature has already acted on potentially synonymous sites in the mitochondrial genome (Figure 4A), meaning that any new somatic changes are much less likely to be silent. In keeping with this, a dN/dS ratio (Materials and Methods) calculated taking into account both the mutational signature and the mtDNA codon usage revealed that missense mutations accumulate at a frequency very close to that expected under neutrality (dN/dS=1.21; 95% confidence interval, 1.015 - 1.434; p=0.031). This indicates that despite the apparent high

ratio of missense to silent mutations, the vast majority of mtDNA mutations are passengers with no convincing evidence suggesting the existence of driver mitochondrial DNA mutations. Additional gene-by-gene analysis further revealed that no single gene had a higher than expected rate of missense or nonsense mutations (Supplementary file 4).

For nonsense substitutions and frameshift indels, we observe a somewhat different picture. Taking into account the mutation signature and amino acid composition of the mitochondrial genome, the overall ratio of nonsense mutations to silent mutations is exactly that expected by chance (dNonsense/dS=1.004; 95% confidence interval, 0.699-1.443; p=0.98). However, while missense and silent substitutions exhibited equivalent variant allele fractions (average VAFs; 40.1% and 40.9%, respectively; p=0.8), nonsense substitutions presented significantly lower VAFs (average 26.4%; $p=6x10^{-5}$), as did frameshift indels (average 25.0%; $p=2x10^{-3}$; Figure 5A). Taken together, these data suggest that nonsense mutations occur at the expected rate given the underlying mutational process. However, while silent and missense substitutions frequently achieve high allele fractions in tumor cells due to the effects of random drift, there are significantly greater constraints on mitochondrial genomes carrying protein-inactivating mutations. The inference here is that cancer cells carrying such deleterious mutations at or near homoplasmy are at a selective disadvantage, and hence do not contribute to clonal expansions, underlining the importance of functional mitochondria to cancer cells. The extent of such disadvantage may vary according to tumor type: for example colorectal cancers show less negative selection compared to breast cancers (p=0.028; Figure 5-figure supplement 2).

We found 171 mtDNA substitutions in mitochondrial tRNA sequences, which is very similar to the expected number (168.2, p=0.82) from the mutational signature. Interestingly, none of the substitutions was located in the trinucleotide anticodon site of the tRNA (expected number=7.6, p=0.006). This suggests mutations in tRNA anticodons confer a similar selective disadvantage as protein-truncating mutations, presumably because such mutations would lead to systematic erroneous aminoacylation of nascent proteins during translation of the relevant codon.

Next, we assessed whether any specific somatic mutations showed evidence of positive selection. Out of the 1,907 somatic substitutions, 16 (0.8%) overlapped with known disease-associated mtDNA mutations, such as mutations frequently detected in MELAS (Mitochondrial Encephalomyopathy, lactic acidosis, and stroke-like episodes) and LHON (Leber hereditary optic neuropathy) (Supplementary file 2). In addition, ten mutations within mitochondrial protein-coding, tRNA and rRNA genes showed significantly higher recurrent rate than expected from background mutational signature (Supplementary file 5). However, it remains unclear whether this high recurrence reflects positive selection, because any factors not included in our background model of the mutational process, such as local mutation hotspots, could also explain a mild excess of mutations at a given nucleotide.

mtDNA Mutations across Tumor Evolution

We investigated whether somatic mtDNA mutations are more likely to become homoplasmic later in tumor evolution by assessing paired cancer samples, either primary and metastasis (breast, colorectal and prostate) or primary and relapse (myeloma) (Figure 5B and

Supplementary file 1). As mentioned earlier, 73 late (metastasis or relapse) cancer samples were sequenced in addition to the primary tissues. Among the mtDNA mutations identified in either of the paired cancer samples, a number of different patterns were observed. There were mutations at high VAF in the primary not found in the metastasis (n=49); mutations in the metastasis not found in the primary (n=49); and shared mutations (n=71) at high or low VAF, sometimes with evidence for drift (VAF difference > 0.2) between the two samples (n=25). These data, particularly the mutations found in the metastasis only, suggest that mitochondrial mutations can occur throughout the time course of tumor evolution, and still drift to homoplasmy with appreciable frequency, as suggested previously (Coller et al., 2001). To assess the plausibility of this conclusion, we modeled the dynamics of mtDNA mutations based on a few simplifying assumptions (Materials and Methods, Evolutionary dynamics of neutral mitochondrial mutations). We find that the expected number of neutral mitochondrial mutations drifting to homoplasmy increases linearly with mutation rate and number of cell divisions. Based on a mutation rate of 10^{-7}/base-pair/generation (Coller et al., 2001; Hudson and Chinnery, 2006), this leads to an average ~1 homoplasmic mutation for every 1,000 cell generations.

Origins of mtDNA somatic mutations

We also explored whether the mutational forces that are so critical to shaping the nuclear genome during tumor evolution could affect the mitochondrial genome. In cancers associated with exogenous mutagens, such as tobacco-associated lung cancer and ultraviolet light-associated melanomas, we found no evidence of the mutational signatures characteristic of these carcinogens among the mtDNA mutations (Figure 5C, Figure 5-figure supplement 3). Moreover, *BRCA1* and *BRCA2* mutations showed no evident influence on mitochondrial genomes in breast

cancer (Figure 5C), in contrast to their effects on nuclear genomes exhibiting an even distribution of mutations across all trinucleotide contexts (Alexandrov et al., 2013; Nik-Zainal et al., 2012a). Taken together, it appears that the primary mtDNA mutational process is endogenous to mitochondria and is very different to those operating in nuclear DNA. It is surprising that the endogenous mutational process has far greater impact than any external forces, as the physicochemical interactions of ultraviolet light or the chemicals in cigarette smoke with DNA should be similar in both genomes. The simulations described above suggest the major explanation to be that the endogenous mutation rate is several orders of magnitude greater than that expected for exogenous carcinogens, thus swamping any signal.

DISCUSSION

In theory, there are two potential sources of the mtDNA variants we observed in cancer tissues: (1) somatically acquired, or *de novo*, mutations accumulated during the cancer clone's lineage of cell divisions from the fertilized egg or (2) low-level heteroplasmic mtDNA present in the oocyte (therefore maternally inherited) amplified in cancer but lost from normal tissue by random drift (Freyer et al., 2012; He et al., 2010; Payne et al., 2013). We believe the majority of the variants we find are genuinely acquired somatically. First, of the 45 pairs of somatic mutations phased together on the same copy of the mtDNA genome, at least 33 (73.3%) showed a clear subclonal relationship and therefore their occurrence is separated in time, or apparently somatic. Secondly, 63.4% of our substitutions were not previously reported as germline polymorphisms. This is a much higher rate than reported for equivalent analyses on heteroplasmic variants in non-cancer samples (8/37; 21.6%) (Li et al., 2010), although methodological differences may somewhat contribute to this apparent difference (Avital et al., 2012; Goto et al., 2011). Thirdly, if the variants were due to inherited, low-level heteroplasmy, we would not expect to see such variation across tissue types, since all tissue types derive from the fertilized egg. It is difficult to distinguish whether the variants we observe occur before or after the initiating driver mutations that herald tumorigenesis, but our analysis of paired samples does suggest that they can occur both early and late. Given the homogeneity of the mutational signature across tumor types and its inferred resemblance to the germline mtDNA mutational process, we would hypothesise that new mutations occur at a fairly constant and high rate per mitochondrial genome replication throughout all cell divisions.

On the basis of the mutational signature observed here, somatic substitutions are unlikely to be attributable to reactive oxygen species (ROS), as previous reports have suggested (Larman et al., 2012; Polyak et al., 1998). Guanine oxidation by ROS predominantly causes G:C>T:A transversion (Delaney et al., 2012; Thilly, 2003), which constitute only 4.0% of the mutations in our data (Figure 5C). Instead, we propose three replication-coupled mechanisms that can explain the strand asymmetric $C_H>T_H$ and $T_L>C_L$ mutational signature and define a model of the mtDNA mutational process (Figure 5D). First, the parent H strand, displaced and single-stranded during mtDNA replication (Holt and Reyes, 2012), could be more prone to cytosine deamination (generating $C_H>T_H$) and/or adenine deamination (Faith and Pollock, 2003; Lindahl, 1993; Saccone et al., 1999) (generating $T_L>C_L$). Secondly, endogenous mtDNA polymerase (*POLG*) replication errors (Zheng et al., 2006) (which show the pattern of C>T and A>G substitutions) could be preferentially generated on the leading strand (Pavlov et al., 2002). Thirdly, there may be differences between the efficiency of repair between the leading and lagging strand (Pavlov et al., 2003). Further, the mutation pattern reported here is consistent with the hypothesized bidirectional initiation of mtDNA genome replication (Holt and Reyes, 2012; Yasukawa et al., 2006; Yasukawa et al., 2005).

It appears that most of the mtDNA missense mutations we observe become fixed in tumor progenitor cells without distinct physiological advantage. All the statistical testing performed in this study – variant allele fraction comparison across different categories of somatic mutations, number of recurrent mutations and dN/dS ratio – suggest that mtDNA somatic substitutions accumulate largely neutrally. This is not different from previous observations in nuclear genomes: of the thousands of somatic mutations found in a cancer genome, many fewer than a

hundred are believed to confer a selective advantage to the cancer cell (Stratton et al., 2009). In contrast, protein-truncating mutations showed evidence of negative selection, at the level of constraints on the allele fraction achieved. The implication of this is that the inactivating mutations occur at an appreciable rate, but the fraction of mitochondrial genomes per cell carrying these variants cannot increase beyond a certain limit without impairing the selective fitness of that cell. Having a sizable number of mitochondria with fully intact proteome remains critical to the fitness of a cancer cell.

MATERIALS AND METHODS

Sequencing data

All the sequences were generated by Illumina platforms (either Genome Analyzer or HiSeq 2000). With respect to TCGA data, we downloaded aligned bam files through UCSC CGHub (http://cghub.ucsc.edu). Sequencing reads were aligned on the human reference genome build 37 (GRCh37) and human reference mtDNA sequence (revised Cambridge reference sequence, rCRS (Andrews et al., 1999)), mainly by BWA alignment tool. Samtools (Li and Durbin, 2009) and Varscan2 (Koboldt et al., 2012) were used for manipulating sequence reads and for calling somatic mutations, respectively. Sequence data have been deposited in the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/home ; study accession # EGAS00001000968; dataset accession numbers EGAD00001001014 for primary samples and EGAD00001001015 for metastatic samples). Sample accession numbers are available in the Supplementary file 6.

Off-target mtDNA reads in whole-exome sequencing

Most of the currently available whole-exome capture kits, including Agilent Technologies SureSelect Human All Exon 50Mb (Agilent Technologies Inc.) used mostly in this study, do not target mtDNA genes (Falk et al., 2012). However, because of the abundance of mtDNA in human cells (100-100,000 copies per cell), it is expected that a number of mtDNA fragments could be off-target captured. We checked whether the amount of off-target mtDNA reads was sufficient for mtDNA variant detection. Whole-exome sequencing (normal samples) generated by CGP (n=855), WUGSC (Washington University Genome Sequencing Center; n=140) and BCM (Baylor College of Medicine; n=85) contained ~100 off-target mtDNA reads per 1M autosomal reads (Figure 1-figure supplement 4). We concluded these could be sufficient for the downstream analyses, because ordinary 10Gb whole-exome data would provide ~60x read-depth for mtDNA here. However, whole-exome data sequenced by BI (Broad Institute; n=436) included far less, ~3 off-target mtDNA per 1M autosomal reads, which would show ~2x mtDNA read-depth per 10Gb exome sequencing (Figure 1-figure supplement 4). It may be due to “improved” exome-capture protocols by BI to increase the DNA-capture efficiency and on-target rate (Fisher et al., 2011). Therefore, we did not include whole-exome data sequenced from BI for further analysis.

139 samples were sequenced by both whole-genome and whole-exome sequencing. From these, we compared the amount of off-target mtDNA reads from whole-exome sequencing with that of whole-genome sequencing. It showed clear positive linear correlation (Figure 1-figure supplement 1).

DNA cross-contamination

Given the abundance of mtDNA in the cancer cells, 1-214x coverage cancer whole nuclear genome sequencing provides extensive coverage of mtDNA (average read-depth = 7901.0x; table S2) enabling accurate identification of somatic mutations, even if heteroplasmic. Whole exome sequencing data were also included because off-target reads provided sufficient coverage (average read-depth = 92.1x) to analyze mtDNA mutations.

This high coverage of mtDNA, especially from whole-genome sequencing, permitted us to identify heteroplasmic variants (our detection threshold was 3%; see "Variant calling" for more details). However, because sample-swaps and/or DNA cross-contaminations would definitely generate false-positive somatic variants, we filtered out suspicious DNA samples as described below.

1) Major sample-swaps

A subset of tumor and normal sequencing pairs, of which the nuclear genotypes were not matching with each other, were removed from further analyses. We randomly selected 320 common single-nucleotide polymorphism sites on the 22 human autosomes, of which the minor allele frequency is ~50% (45-55%) according to The 1000 Genomes Project (Genomes Project et al., 2010). Of the 320 sites, homozygous positions in normal tissues (where showed >90% variant allele fraction (VAF) with bases Q score>20) were compared with the corresponding genotypes in the counterpart cancer. Sample pairs were removed if the genotype mismatch rate was greater than 0.1 ($\frac{Nhet+Nwt}{Nhom+Nhet+Nwt}$; N_{het}, number of heterozygote positions; N_{hom}, number of homozygote positions; N_{wt}, number of wildtype positions)(Figure 1-figure supplement 5A). We note 0 is expected for the rate when genotyping is perfect and sample pairs are from the same individual. By contrast, 0.5 is expected when samples were from different individuals.

2) Minor cross-contamination

We estimated DNA cross-contamination levels with the VAF of autosomal homozygous SNPs genotyped from the common (population minor allele frequency ~50%) SNP sites. Theoretically, if there is no sequencing (and mapping) error, all the homozygote SNP sites in pure samples should present 100% VAFs. However, when samples are contaminated, corresponding VAFs are reduced because the contaminant has only a ~25% of chance of having homozygote SNPs on the same site. Therefore, minor contamination levels (C) of each cancer sequencing data were estimated as below:

$$C = 2 \times \frac{\sum(RCwt) - Ne}{\sum(RDhom) - Ne}$$

, where RD_{hom} is sequencing read-depth, RC_{wt} is readcount of wildtype alleles and Ne is number of sequencing errors on each autosomal homozygote SNP site. For high accuracy, we only counted base with sufficient quality score (Q>20). In order to estimate Ne, we assumed a conservative rate (sequencing error rate = 0.001). We considered sites covered by at least 10 reads and 90% VAF (Figure 1-figure supplement 5B). 95% confidence intervals of cross-contamination levels were calculated using binomial distribution.

In order to clear somatic variants, here we made the very conservative assumption that somatic variants present in excess of 5-times of the 95% upper limit of C levels were true somatic rather than false positives by low-level of cross-contamination.

3) Germline polymorphisms and back mutations

We further checked samples for contamination using known mtDNA polymorphisms. Because human mtDNA is small (16,569 bp) and extensively explored previously, most of germline mtDNA polymorphisms are already known. For example, 97.7% of the 39,036 inherited substitutions were known polymorphisms in the mtDB database (Ingman and Gyllensten, 2006). Therefore, when a tumor sample is contaminated by other samples, many somatic-like mtDNA substitutions by contaminants are likely to be overlapped with known mtDNA polymorphisms.

At the same time, low-level contamination would generate excessive back mutations, which appeared to reverse germline common polymorphisms into wildtype alleles. Taken together, both the number of somatic substitutions known in mtDB and number of back mutations can be good indicators for mtDNA cross-contamination. Therefore, we filtered out tumor tissues with ≥ 3 known potentially somatic mutations or with ≥ 2 back mutations from the further analyses (Figure 1-figure supplement 5A and B).

Variant calling

We extracted mtDNA reads using Samtools (Li and Durbin, 2009). We used VarScan2 (Koboldt et al., 2012) for initial variant calling with a few options (--strand-filter 1 (mismatches should be reported by both forward and reverse reads), --min-var-freq 0.03 (minimum VAF 3%), --min-avg-qual 20 (minimum base quality 20), min-coverage 3 and --min-reads2 2). With respect to the --strand-filter, it generally removes variant when >90% of mismatches are reported from either of the H or the L mtDNA strand. However, where only reads with a specific orientation are could be aligned dominantly (i.e. in both extreme region of mitochondrial reference genome; only L strand reads could be aligned on the 5' extreme of mtDNA), we compared strand bias between "perfect matches" (# perfect matches from L strand reads / total # perfect matches) and mismatches (# mismatches from L strand reads / total # mismatches). If the difference between those two bias < 0.1, the mutations were rescued. Of the 1,907 mutations, 54 (2.8%) were rescued accordingly.

Putative somatic variants called by VarScan2 were further filtered using criteria shown below.

(1) At least 4 unique reads supporting variants AND all variant reads at least 20 phred-scale sequencing quality score (Q 20 = 1% sequencing error rate) AND at least 3% variant allele fractions (VAFs).
    a. Regardless of in WGS and in WES, the ≥4 mismatches and the ≥3% VAF criteria must be satisfied simultaneously.
    b. However, in WGS, the minimum number of reads (n=4) criterion is not essential, because the ≥3% VAF criterion is much more stringent (3% VAF request at least 240 mismatches (>>4) given mtDNA coverage is ~8,000 for WGS).
    c. In WES, the ≥3% VAF criterion is relatively less important than in WGS, because the ≥4 mismatches criterion is more stringent. For example, 4 mismatches in 90x (WXS average)

coverage region (VAF=4.4%) automatically fulfill the ≥3% VAF criterion. For less covered regions (i.e. <40x coverage; n=285 out of total 1,907 substitutions), the VAF criterion becomes less important, because 4 mismatches would generate ≥10% VAF, much higher than the minimum threshold (i.e. 3%). As results, we are missing lower heteroplasmic variants (i.e. variants with 3%-10% heteroplasmic levels) from low coverage samples (mostly by WXS). The lower sensitivity of WXS is also confirmed in our validation study (See "Validation of somatic variants" below).

(2) There is no minimum threshold for total coverage (# perfect matches + # mismatches).

(3) To increase sensitivity for detecting mutations, we rescued mutations with 3 unique variant reads (with at least 20 phred-scale sequencing quality score) when VAFs is ≥20%. Of 1,907 somatic substitutions, 32 (1.7%) were rescued accordingly.

(4) All somatic variants presenting with VAFs lower than our very conservative threshold for minor cross-contamination (5-times 95% upper limit of contamination levels for each tumor sample, see above "Minor cross-contamination of DNA samples") were removed. When we could not estimate cross-contamination levels because of low sequencing depth of coverage (for nuclear genome), a conservative criterion (10% contamination level threshold) was explicitly used.

(5) Substitutions were further visually inspected using IGV (Thorvaldsdottir et al., 2013). Thirteen frequent false positive variants (shown below) by misalignment due to extensive level of homopolymers in rCRS and due to sequencing error in the reference mtDNA genome (3107N, see Mitomap (http://www.mitomap.org/bin/view.pl/MITOMAP/CambridgeReanalysis) for more information) were explicitly removed:

1) misalignment due to ACCCCCCCTCCCCC (rCRS 302-315)
A302C, C309T, C311T, C312T, C313T, G316C
2) misalignment due to GCACACACACACC (rCRS 513-525)
C514A, A515G, A523C, C524G
3) misalignment due to 3107N in rCRS (ACNTT, rCRS 3105-3109)
C3106A, T3109C, C3110A

We compared our variant calls with common inherited mtDNA polymorphisms deposited in the mtDB database as of 24. Jul. 2013 (Ingman and Gyllensten, 2006). Gene annotation of somatic variants was done using custom script based on human mtDNA gene information (Ruiz-Pesini et al., 2007).

Validation of somatic variants

To validate the sensitivity and specificity of variant calling in this study, 19 tumor and normal pairs (which were originally whole-genome sequenced) were whole-exome sequenced and mtDNA variants were assessed independently. Among the 28 somatic substitutions originally detected from the 19 tumor-normal whole-genome

sequencing pairs, 20 (71.4%) were called as somatic (Figure 1-figure supplement 3). In addition, 5 (17.9%) presented evidence of variant reads in the validation set, although it was filtered out because of its low read-depth of coverage in exome sequencings (showed 2-5 variant reads). Moreover, because 3 remaining sites were not sufficiently covered in the validation set to call somatic variants, these could not be evidence of the inaccuracy of whole-genome sequencing data, therefore not considered in the accuracy validation. Taken together, all the 25 somatic substitutions by whole-genome sequencing were highly likely to be true positives, therefore we concluded it provided ~100% accuracy in the mtDNA somatic substitution assessment. Actually, the high accuracy of whole-genome sequencing is very likely and what we expect, because it provides extensive coverage of mtDNA (average read-depth > 7,500x), ~3% heteroplasmic variants would present >200 variant reads.

By contrast, the validation set (whole-exome sequencing) called 21 somatic substitutions. Of these, 20 were common with whole-genome sequencing, and one was incorrectly called as somatic though it was actually germline substitutions in the whole-genome sequencing data. In addition, as mentioned above, the validation set missed 8 somatic substitutions called by whole-genome sequencing. Six out of eight undercalls (75%) were low heteroplasmic substitutions in whole-genome sequencing, ranging from 3.36% to 8.68%. Based on these data, we suggest 71.4% sensitivity (20/28) and 95.2% specificity (20/21) for exome-sequencing in detecting upto 3% heteroplasmic somatic mtDNA substitutions in cancer.

We further checked the correlation of heteroplasmy level between the 20 mtDNA somatic mutations called both whole-genome and whole-exome sequencing. It showed great linear relationship (R^2=0.97, Figure 1-figure supplement 2), further suggesting whole-exome sequencing data is appropriate for accurate detection of mtDNA somatic mutations.

Substitution phasing

We phased 72 somatic substitution pairs, which arose in a single cancer sample and which located sufficiently close (from 10bp to ~500bp) therefore both sites could be sequenced by same sequence fragments (Supplementary file 3 and Figure 2-figure supplement 1). We classified them as ‘different strand’, ‘co-clonal’ and ‘sub-clonal’ using criteria as follow:

Different strand: the two somatic substitutions are obligate on different strands. Reads that report wildtype1(wt)-substitution2(subs) and subs1-wt2, but subs1-subs2, are observed.

Co-clonal: reads reporting wt1-wt2 and subs1-subs2 are only observed.

Subclonal: One substitution is subclonal to the other, but the two are definitely phased. Reads subs1-subs2 and either subs1-wt2 or wt1-subs2 are observed.

Tumor type and mtDNA somatic substitutions

To understand the relationship between tumor types and number of mtDNA mutations, Poisson regression and ANOVA was applied to our dataset using R software (http://www.r-project.org).

Fit1 <- glm(N_{sub} ~ Cov_T + Cov_N, family=poisson())

Fit2 <- glm(N_{sub} ~ Cov_T + Cov_N + t, family=poisson())

anova(Fit1, Fit2, test="Chisq")

, where N_{sub} is number of mtDNA substitutions of each sample, Cov_T and Cov_N are coverage of tumor and normal mtDNA, respectively, (if Cov is >200, we replaced it by 200), t is tumor types.

Age and mtDNA somatic substitutions

Poisson regression was applied to our breast cancer dataset.

Fit1 <- glm(N_{sub} ~ Cov_T + Cov_N + a, family=poisson())

, where N_{sub} is number of mtDNA substitutions of each sample, Cov_T and Cov_N are coverage of tumor and normal mtDNA, respectively, (if Cov is >200, we replaced it by 200), a is age at diagnosis. P-value in estimation of a was shown in the manuscript.

Mutational signature and strand bias

Different mutational processes generate different combinations of mutation types, termed "signatures" (Nik-Zainal et al., 2012a). For example, ultraviolet (UV) light and tobacco smoking (polycyclic aromatic hydrocarbons) frequently generate C>T transitions and G>T transversions on non-transcribed (coding) strands in melanoma and lung cancers, respectively (Pleasance et al., 2010a; Pleasance et al., 2010b). To understand the mutational processes influencing cancer mtDNA, we correlated the 1,907 mtDNA substitutions with 21 cancer specific mutational signatures in the nuclear DNA recently identified (Alexandrov et al., 2013). However, none of the signature could explain the highly unique mtDNA substitutions.

Mutational signature and strand bias was assessed as described in our previous reports (Alexandrov et al., 2013). Briefly, The immediate 5′ and 3′ sequence context was extracted from rCRS. Substitution rate for each trinucleotide context was calculated with the number of substitution normalized by the frequency of the trinucleotide context observed in the rCRS, in the L and H strand, respectively. For analyses of substitutions falling in the mtDNA genes (13 protein-coding and 22 tRNA genes), transcribed/non-transcribed strand was also considered for comparison.

In order to prove the strand bias is not transcription but replication-coupled, we checked strand biases of polymorphisms in the 12 L strand protein-coding genes, 1 H strand protein-coding gene (*MT-ND6*) and/or 22 tRNAs (Figure 3 – supplement1). For this specific purpose, we did not consider the sequence context (immediate 5' and 3' bases) because it over-classifies mutations (i.e. the number of mutation classes (n=96) is larger than that of mutations). In other words, 12 classes of substitutions (six classes of possible base substitutions (C>A, C>G, C>T, T>A, T>C, T>G) x two strands (L and H strands)) were considered. Substitution rates are ratio between observed and expected numbers (H_0=same mutation rate for all substitution classes) for each substitution class. In order to

understand which model (replicative or transcriptional strand) is appropriate to explain the strand-bias, chi-square tests were used between numbers of observed mutations for each class and expected ones under the background signature.

mtDNA codon usage

We counted the codon frequencies in 13 mtDNA protein-coding genes. Because 12 L strand protein-coding genes and 1 H strand gene (*MT-ND6*) are under opposite mutational pressure (T>C and G>A for L strand genes; A>G and C>T for *MT-ND6*), we separated L and H strand genes for this analysis. T>C skew and G>A skew were calculated as shown below, to understand the $T_L>C_L$ and $C_H>T_H$ (equivalent to $G_L>A_L$) substitutions during the evolution of human mtDNA:

$$T > C_{skew} = \frac{N_C - N_T}{N_C + N_T} \quad and \quad G > A_{skew} = \frac{N_A - N_G}{N_A + N_G}$$

, where N_A, N_C, N_G and N_T are number of A, C, G, and T base in the 3rd position of triplet codons in mtDNA genes, respectively.

For the assessment of mtDNA codon usage of other animal species, we analyzed the mtDNA sequence of *C. elegans* (accession# NC_001328), *D. melanogaster* (accession# NC_001709), *D. rerio* (accession# NC_002333), *X. laevis* (accession# NC_001573), *M. musculus* (accession# EU450583), *G. Domesticus* (accession # NC_235570), and *P. troglodytes* (NC_001643). We considered only L strand mtDNA genes in the cross-species analysis.

Recurrent substitutions

To compare the number of recurrent substitutions between silent and missense substitutions, we randomly selected 100 substitutions each from 198 silent substitutions in the 3rd base of triplet codons, 440 missense substitutions in the 1st base of triplet codons, and 405 missense substitutions in the 2nd base of triplet codons. We counted numbers of recurrent substitutions in each group. This was iterated 300 times independently. ANOVA testing was applied to determine the difference between the three groups (Figure 5-figure supplement 1).

dN/dS ratio

To estimate dN/dS values for missense mutations (w_{mis}), we used an adaptation of the method described previously (Greenman et al., 2006). Briefly, the rate of mutations is modeled as a Poisson process, with a rate given by a product of the mutation rate and the impact of selection. To obtain accurate estimates of dN/dS we used two separate models, one using 12 single-nucleotide substitution rates and a more complex one accounting for any context-dependence effect by 1-nucleotide upstream and downstream using 192 substitution rates. For example in the 12-rate model, the expected number of A>C mutations ($\boldsymbol{\lambda}_{A>C}$) would is modeled as

$\boldsymbol{\lambda}_{syn,A>C} = r_{A>C} * L_{syn,A>C}$

$\boldsymbol{\lambda}_{mis,A>C} = r_{A>C} * w_{mis} * L_{mis,A>C}$

, being $L_{syn,A>C}$ and $L_{mis,A>C}$ the number of sites that can suffer a synonymous and missense A>C mutation, respectively, which are calculated for any particular sequence. The likelihood of observing the number of missense A>C mutations ($N_{mis,A>C}$) given the expected $\boldsymbol{\lambda}_{mis,A>C}$ is then calculated as:

$$Lik = Poisson(N_{mis,A>C} \mid r_{A>C}, w_{mis})$$

and the likelihood of the entire model is the product of all individual likelihoods. W_{mis} is fixed to be equal in all 12 (or 192) equations describing each substitution type and a hill-climbing algorithm is used to find the maximum-likelihood estimates for all rate and selection parameters. Likelihood Ratio Tests are then used to test deviations from neutrality ($w_{mis} = 1$). The dN/dS ratio reported in the main text corresponds to the full context-dependent model with 192 substitution rates. This method allows quantifying the strength of selection avoiding the confounding effect of gene length, sequence composition, different rates of each substitution type and context-dependent mutagenesis.

Short indels

Along with the 1,907 somatic mtDNA substitutions, we identified 109 and 142 somatic short insertions and deletions, respectively, from the 1,675 cancer mtDNA sequences using Varscan2 (Supplementary file 2).

Evolutionary dynamics of neutral mitochondrial mutations

We model the evolutionary dynamics of mitochondrial mutations under random drift and derive a simple equation for the expected number of homoplasmic mutations. There exist multiple levels at which mitochondrial mutations evolve: within mitochondria, in the cytoplasm and on the cellular level (Rand, 2011). Here we focus on the dynamics in a single cell, which represents the founder of the last clonal expansion in the tumor cell population. The cellular dynamics during a clonal expansion are difficult to describe analytically, but it is important to realize that mutations a clonal expansion preserves the allele frequencies of neutral variants and that mutations that occur after the expansion are unlikely to contribute to measurable allele frequencies, as the population becomes large.

We model the evolutionary dynamics of mitochondrial mutations in the cytoplasm of a single cell by a Wright-Fisher process (Wright, 1931), in which the number of mitochondria in a subsequent generation is a binomial sample of the mitochondria in the previous generation. The number of mitochondria M is kept fixed. The marginal allele frequency X of a single site has two absorbing boundaries X=0 and X=M (homoplasmy) and the probability of fixation of an allele at frequency X by neutral drift is $\rho = X/M$ (Wright, 1931). Note that this process leads, on the population level, to a dichotomization of heteroplasmic variants to either go extinct or become homoplasmic and fixate in a cell.

Mutations on any of L (=16,569nt) sites in the mitochondrial genome are assumed to occur at a uniform rate μ per nucleotide per cell division, which is of order 10^{-7}, based on a human inter-generational comparison (Coller et al.,

2001). Hence the rate of neutral evolution is simply $\mu LM / M = \mu L$ (Kimura, 1984). Lastly, the expected time to fixation in the Wright-Fisher process is $t = 2M$. Putting these things together, the expected number of mutant alleles N in a cell initially without any mitochondrial mutations after T generation is

$$E[N] = \mu L (T - 2M)$$

This equation predicts a linear accumulation of neutral mutations over time, with a delay imposed by number of mitochondrial copies. A similar behavior has been reported using numerical simulations (Coller et al., 2001). When also considering heteroplasmic mutations, the expected number of alterations may be slightly higher.

To check whether our model yields the correct behaviour, we use the following numbers: The observed order of magnitude of mitochondrial mutations per patient was N=1. The sequencing coverage on the mitochondrial genome indicates that there we of order M=100 mitochondrial genome copies present per cancer cell. The expected number of mutations per cell division is $\mu L = 1.6 \times 10^{-3}$, it therefore requires around 1000 cell generations T to accumulate on average one homoplasmic mutation. This number of generations appears realistic for regenerating tissues. As expected, epithelial cancers had among the highest observed number of mitochondrial mutations, while hematopoietic cancers typically had lower numbers.

Statistical testing

Statistical testing was performed using R software. All p-values were calculated by two-tailed testing. Figures were generated using R and Microsoft Excel software.

ACKNOWLEDGEMENTS

Data used in this manuscript are described in the supplementary materials (Supplementary file 1). We thank Thomas Bleazard at Faculty of Medical and Human Sciences, University of Manchester for discussion and assistance with manuscript preparation. We would like to thank The Cancer Genome Atlas (TCGA) Project Team and their specimen donors for providing sequencing data. This work was supported by the Wellcome Trust, the British Lung Foundation, the Health Innovation Challenge Fund, the Kay Kendall Leukaemia Fund, the Chordoma Foundation and the Adenoid Cystic Carcinoma Research Foundation. Y.S.J and I.M. are supported by EMBO long-term fellowship (ALTF 1203-2012 and ALTF 1287-2012, respectively). P.J.C. is a Wellcome Trust Senior Clinical Fellow. Support was provided to A.M.F. by the National Institute for Health Research (NIHR) UCLH Biomedical Research Centre. A.R.G. receives support from Leukaemia Lymphoma Research, Cancer Research UK and the Leukemia Lymphoma Society. Samples from Addenbrooke's Hospital were collected with support from the NIHR Cambridge Biomedical Resource Centre. The ICGC Breast Cancer Consortium was supported by a grant from the European Union (BASIS) and the Wellcome Trust. The ICGC Prostate Cancer Consortium was funded by Cancer Research UK. We would also like to acknowledge the support of the National Cancer Research Prostate Cancer: Mechanisms of Progression and Treatment (PROMPT) collaborative (grant code G0500966/75466) which has funded tissue and urine collections in Cambridge. This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (J.A.T.). We obtained informed consent and consent to publish from participants enrolled.

COMPETING INTERESTS

The authors declare that they have no conflict of interest with this manuscript.

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L*., et al.* (2013). Signatures of mutational processes in human cancer. Nature.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature genetics *23*, 147.

Avital, G., Buchshtav, M., Zhidkov, I., Tuval Feder, J., Dadon, S., Rubin, E., Glass, D., Spector, T.D., and Mishmar, D. (2012). Mitochondrial DNA heteroplasmy in diabetes and normal adults: role of acquired and inherited mutational patterns in twins. Human molecular genetics *21*, 4214-4224.

Brandon, M., Baldi, P., and Wallace, D.C. (2006). Mitochondrial mutations in cancer. Oncogene *25*, 4647-4662.

Calvo, S.E., and Mootha, V.K. (2010). The mitochondrial proteome and human disease. Annual review of genomics and human genetics *11*, 25-44.

Chatterjee, A., Mambo, E., and Sidransky, D. (2006). Mitochondrial DNA mutations in human cancer. Oncogene *25*, 4663-4674.

Chinnery, P.F. (1993). Mitochondrial Disorders Overview. In GeneReviews, R.A. Pagon, M.P. Adam, T.D. Bird, C.R. Dolan, C.T. Fong, R.J.H. Smith, and K. Stephens, eds. (Seattle (WA)).

Clayton, D.A. (1982). Replication of animal mitochondrial DNA. Cell *28*, 693-705.

Coller, H.A., Khrapko, K., Bodyak, N.D., Nekhaeva, E., Herrero-Jimenez, P., and Thilly, W.G. (2001). High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. Nature genetics *28*, 147-150.

De Alwis, N., Hudson, G., Burt, A.D., Day, C.P., and Chinnery, P.F. (2009). Human liver stem cells originate from the canals of Hering. Hepatology *50*, 992-993.

Delaney, S., Jarem, D.A., Volle, C.B., and Yennie, C.J. (2012). Chemical and biological consequences of oxidatively damaged guanine in DNA. Free radical research *46*, 420-441.

Ericson, N.G., Kulawiec, M., Vermulst, M., Sheahan, K., O'Sullivan, J., Salk, J.J., and Bielas, J.H. (2012). Decreased mitochondrial DNA mutagenesis in human colorectal cancer. PLoS genetics *8*, e1002689.

Faith, J.J., and Pollock, D.D. (2003). Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics *165*, 735-745.

Falk, M.J., Pierce, E.A., Consugar, M., Xie, M.H., Guadalupe, M., Hardy, O., Rappaport, E.F., Wallace, D.C., LeProust, E., and Gai, X. (2012). Mitochondrial disease genetic diagnostics: optimized whole-exome analysis for all MitoCarta nuclear genes and the mitochondrial genome. Discovery medicine *14*, 389-399.

Falkenberg, M., Larsson, N.G., and Gustafsson, C.M. (2007). DNA replication and transcription in mammalian mitochondria. Annual review of biochemistry *76*, 679-699.

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L*., et al.* (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol *12*.

Freyer, C., Cree, L.M., Mourier, A., Stewart, J.B., Koolmeister, C., Milenkovic, D., Wai, T., Floros, V.I., Hagstrom, E., Chatzidaki, E.E*., et al.* (2012). Variation in germline mtDNA heteroplasmy is determined prenatally but modified during subsequent transmission. Nature genetics *44*, 1282-1285.

Genomes Project, C., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

Goto, H., Dickins, B., Afgan, E., Paul, I.M., Taylor, J., Makova, K.D., and Nekrutenko, A. (2011). Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biol *12*, R59.

Gray, M.W., Burger, G., and Lang, B.F. (1999). Mitochondrial evolution. Science *283*, 1476-1481.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C*., et al.* (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153-158.

Greenman, C., Wooster, R., Futreal, P.A., Stratton, M.R., and Easton, D.F. (2006). Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics *173*, 2187-2198.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

He, Y., Wu, J., Dressman, D.C., Iacobuzio-Donahue, C., Markowitz, S.D., Velculescu, V.E., Diaz, L.A., Jr., Kinzler, K.W., Vogelstein, B., and Papadopoulos, N. (2010). Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature *464*, 610-614.

Holt, I.J., and Reyes, A. (2012). Human mitochondrial DNA replication. Cold Spring Harbor perspectives in biology *4*.

Hudson, G., and Chinnery, P.F. (2006). Mitochondrial DNA polymerase-gamma and human disease. Human molecular genetics *15 Spec No 2*, R244-252.

Ingman, M., and Gyllensten, U. (2006). mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic acids research *34*, D749-751.

Kimura, M. (1984). The neutral theory of molecular evolution (Cambridge University Press).

Koboldt, D.C., Zhang, Q.Y., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res *22*, 568-576.

Koppenol, W.H., Bounds, P.L., and Dang, C.V. (2011). Otto Warburg's contributions to current concepts of cancer metabolism. Nature reviews Cancer *11*, 325-337.

Larman, T.C., DePalma, S.R., Hadjipanayis, A.G., Cancer Genome Atlas Research, N., Protopopov, A., Zhang, J., Gabriel, S.B., Chin, L., Seidman, C.E., Kucherlapati, R.*, et al.* (2012). Spectrum of somatic mitochondrial mutations in five cancers. Proceedings of the National Academy of Sciences of the United States of America *109*, 14087-14091.

Legros, F., Malka, F., Frachon, P., Lombes, A., and Rojo, M. (2004). Organization and dynamics of human mitochondrial DNA. Journal of cell science *117*, 2653-2662.

Levin, L., Zhidkov, I., Gurman, Y., Hawlena, H., and Mishmar, D. (2013). Functional recurrent mutations in the human mitochondrial phylogeny: dual roles in evolution and disease. Genome biology and evolution *5*, 876-890.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, M., Schonberg, A., Schaefer, M., Schroeder, R., Nasidze, I., and Stoneking, M. (2010). Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. American journal of human genetics *87*, 237-249.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature *362*, 709-715.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A.*, et al.* (2012a). Mutational processes molding the genomes of 21 breast cancers. Cell *149*, 979-993.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M.*, et al.* (2012b). The life history of 21 breast cancers. Cell *149*, 994-1007.

Nikolaou, C., and Almirantis, Y. (2006). Deviations from Chargaff's second parity rule in organellar DNA Insights into the evolution of organellar genomes. Gene *381*, 34-41.

Pavlov, Y.I., Mian, I.M., and Kunkel, T.A. (2003). Evidence for preferential mismatch repair of lagging strand DNA replication errors in yeast. Current biology : CB *13*, 744-748.

Pavlov, Y.I., Newlon, C.S., and Kunkel, T.A. (2002). Yeast origins establish a strand bias for replicational mutagenesis. Molecular cell *10*, 207-213.

Payne, B.A., Wilson, I.J., Yu-Wai-Man, P., Coxhead, J., Deehan, D., Horvath, R., Taylor, R.W., Samuels, D.C., Santibanez-Koref, M., and Chinnery, P.F. (2013). Universal heteroplasmy of human mitochondrial DNA. Human molecular genetics *22*, 384-390.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R.*, et al.* (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191-196.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C.*, et al.* (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184-190.

Polyak, K., Li, Y., Zhu, H., Lengauer, C., Willson, J.K., Markowitz, S.D., Trush, M.A., Kinzler, K.W., and Vogelstein, B. (1998). Somatic mutations of the mitochondrial genome in human colorectal tumours. Nature genetics *20*, 291-293.

Rand, D.M. (2011). Population genetics of the cytoplasm and the units of selection on mitochondrial DNA in Drosophila melanogaster. Genetica *139*, 685-697.

Ruiz-Pesini, E., Lott, M.T., Procaccio, V., Poole, J.C., Brandon, M.C., Mishmar, D., Yi, C., Kreuziger, J., Baldi, P., and Wallace, D.C. (2007). An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic acids research *35*, D823-D828.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., and Reyes, A. (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene *238*, 195-209.

Schon, E.A., DiMauro, S., and Hirano, M. (2012). Human mitochondrial DNA: roles of inherited and somatic mutations. Nature reviews Genetics *13*, 878-890.
Smeitink, J., van den Heuvel, L., and DiMauro, S. (2001). The genetics and pathology of oxidative phosphorylation. Nature reviews Genetics *2*, 342-352.
Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719-724.
Taylor, R.W., Barron, M.J., Borthwick, G.M., Gospel, A., Chinnery, P.F., Samuels, D.C., Taylor, G.A., Plusa, S.M., Needham, S.J., Greaves, L.C.*, et al.* (2003). Mitochondrial DNA mutations in human colonic crypt stem cells. The Journal of clinical investigation *112*, 1351-1360.
Thilly, W.G. (2003). Have environmental mutagens caused oncomutations in people? Nature genetics *34*, 255-259.
Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform *14*, 178-192.
Wallace, D.C. (2012). Mitochondria and cancer. Nature reviews Cancer *12*, 685-698.
Wright, S. (1931). Evolution in Mendelian Populations. Genetics *16*, 97-159.
Yasukawa, T., Reyes, A., Cluett, T.J., Yang, M.Y., Bowmaker, M., Jacobs, H.T., and Holt, I.J. (2006). Replication of vertebrate mitochondrial DNA entails transient ribonucleotide incorporation throughout the lagging strand. The EMBO journal *25*, 5358-5371.
Yasukawa, T., Yang, M.Y., Jacobs, H.T., and Holt, I.J. (2005). A bidirectional origin of replication maps to the major noncoding region of human mitochondrial DNA. Molecular cell *18*, 651-662.
Zheng, W., Khrapko, K., Coller, H.A., Thilly, W.G., and Copeland, W.C. (2006). Origins of human mitochondrial point mutations as DNA polymerase gamma-mediated errors. Mutation research *599*, 11-20.

FIGURE LEGENDS AND TABLES

Figure 1. Mitochondrial somatic substitutions identified from 1,675 Tumor-Normal pairs. mtDNA genes and intergenic regions are shown. The strand of genes is shown based on mtDNA strand containing equivalent sequences of transcribed RNA. Substitution categories (silent, non-silent (missense and nonsense), non-coding (tRNA and rRNA) and intergenic) are shown by shapes of each substitution. Six classes of substitutions are presented color-coded. The substitutions on the H, and L strand (when six substitutional class were considered) are shown outside and inside of mtDNA genes, respectively. Vertical axes for H and L strand substitutions represent the VAF of each variant.

Figure 2. mtDNA somatic substitutions of human cancer. (A) Number of somatic substitutions in a tumor sample. **(B)** Average number of somatic substitutions per sample across 31 tumor types. **(C)** Age of diagnosis and number of mtDNA somatic substitutions in breast cancers.

Figure 3. Replicative strand bias for mtDNA somatic substitutions. (A) Replicative strand-specific substitution rate (# of observed / # of expected) by 96 trinucleotide context. Substitutions in a specific mtDNA segment (from Ori-b to O_H) are not included, because they present a different substitutional signature. **(B)** Mutational signature across tumor types. Eighteen tumor types, which include at least 25 mtDNA mutations, were shown. **(C)** Inverted substitution signature in the Ori-b - O_H.

Figure 4. Mutational signature similar to processes shaping human mtDNA sequence over evolutionary time **(A)** triplet codon depletion in human mtDNA by equivalent (C_H>T_H and T_L>C_L) mutational pressure. Relative frequency of each triplet codon within synonymous pairs (NNT-NNC or NNA-NNG) is shown by color. The arrows beside the box highlight the T>C (red) and G>A (blue) substitutional pressures on the L strand in germline mtDNA **(B)** Correlation of triplet codon frequencies between from observed and from simulated evolutions of a random sequence mtDNA by the mtDNA somatic mutational signature with constraining mitochondrial protein sequences.

Figure 5. Selection and mutational process for mtDNA somatic substitutions. (A) Truncating mutations (nonsense substitutions and frame-shifting (FS) coding indels) present significantly lower VAF. **(B)** Change of VAF of mtDNA somatic mutation between primary and metastatic (or late) cancer tissues. **(C)** Mutational signature for mtDNA across various tumor types. None of the three highlighted mechanisms or nuclear DNA double-strand breaks repair mechanism (*BRCA*) match with the mtDNA mutational signature. * Only substitutions in protein-coding genes considered. **(D)** A proposed model of mtDNA mutational process.

Figure Supplements Legends

Figure 1-figure supplement 1. Correlation in amount of mtDNA reads between whole-genome and whole-exome sequencing. 139 DNA samples, either from tumors or bloods, sequenced by whole-genome sequencing were additionally sequenced by whole-exome sequencing. We compared the amount of mtDNA reads between whole-genome and whole-exome sequencing. As shown in this figure, we found strong positive correlation. * CGP; Cancer Genome Project, Wellcome Trust Sanger Institute, WUGSC; Washington University Genome Sequencing Center

Figure 1-figure supplement 2. Correlation of heteroplasmy levels between whole-genome and whole-exome sequencing. To validate the sensitivity and specificity of variant calling in this study, 19 tumor and normal pairs (which were originally whole-genome sequenced) were whole-exome sequenced and mtDNA variants were assessed independently. We correlated the heteroplasmic levels of 20 mutations detected in common.

Figure 1-figure supplement 3. Validation of mtDNA somatic substitutions.

Figure 1-figure supplement 4. Amount of off-target mtDNA reads across four sequencing centers. * CGP; Cancer Genome Project, Wellcome Trust Sanger Institute (n = 855), WUGSC;

995 Washington University Genome Sequencing Center (n=140), BCM; Baylor College of Medicine
996 (n=85), BI; Broad Institute (n=435)
997
998 **Figure 1-figure supplement 5. Filtering samples of potential DNA contaminations. (A)** A
999 histogram presenting potential sample-swaps in tumor-sample pairs. **(B)** A histogram presenting
1000 potential minor DNA cross-contamination in tumor samples. Cross-contamination levels were
1001 considered in filtering substitutions (see "Minor cross-contamination of DNA samples" section
1002 in Materials and Methods). **(C)** Histograms showing number of somatic substitutions
1003 overlapping with known inherited polymorphisms and **(D)** number of back mutations.
1004
1005 **Figure 2-figure supplement 1. VAFs of phased somatic mtDNA substitutions.**
1006 This figure presents VAF pairs between co-clonal, subclonal and different strand mtDNA
1007 substitutions. We expect similar VAFs for co-clonal pairs; lower VAF in sub-clonal mutations
1008 compared to clonal ones; and sum of a VAF pair is equal or less than 1.0.
1009
1010 **Figure 3-figure supplement 1. Replicative strand bias observed in mtDNA substitutions. (A)**
1011 Mutational signature of mtDNA somatic substitutions on the 12 L strand genes by replicative
1012 strand (L/H strand). It agrees very well with the background mutational signature. (Chi-square
1013 p=0.99999) **(B)** Mutational signature of mtDNA somatic substitutions on the H strand gene (*MT-*
1014 *ND6*) by replicative strand. It is very close to the background very close to the expected
1015 background signature (Chi-square p=0.027). If we consider signature by transcriptional strand,
1016 the signature difference is very clear (Chi-square $p=1x10^{-21}$). These suggest the strand bias not to
1017 be transcription-coupled, but replication coupled. **(C)** Mutational spectrum of mtDNA somatic
1018 substitutions on the 22 tRNA genes by replicative strand. Again, it agrees very well with the
1019 background mutational signature. (Chi-square p=0.71) **(D)** Mutational spectrum of mtDNA
1020 somatic substitutions on the 22 tRNA genes by non-transcribed (coding) and transcribed (non-
1021 coding) strand. Strand bias was greatly subsided because somatic substitutions on 14 L strand
1022 and 8 H strand tRNAs neutralize the strand bias ($C_H>T_H$ and $T_L>C_L$) each other. As a result, this
1023 signature of tRNA mutations by transcriptional strand is significantly different from the
1024 background one (Chi-square $p=3.3x10^{-12}$). Taken all together, we concluded that the cause of
1025 strand bias is not transcription-coupled but is replicative.
1026

Figure 4-figure supplement 1. TC and GA skew for L strand mtDNA genes across 8 animal species. *C.elegans* (a nematode) and *D. melanogaster* (fruit fly) mtDNA appears to have $G_L<<A_L$ (due to $C_H>T_H$ mutational pressure) and $C_L>>T_L$ (due to $C_L>T_L$ mutational pressure) in the 3rd base of triplet codon in L strand genes. Therefore they seem to have predominant C>T mutational pressure without strand bias. *D. rerio* (zebrafish), *X. laevis* (frog) and *M. musculus* (mouse) presents $G_L<<A_L$ (due to $C_H>T_H$ mutational pressure), but similar number of C_L and T_L. Therefore, mtDNA of these sequences is thought to have $C_H>T_H$, with strand bias. The existence of $T_L>C_L$ is not clear. Finally, mtDNA of *H. sapiens*, *P. troglodytes* (Chimpanzee) and *G. domesticus* (Chicken) shows clear $C_H>T_H$ and $T_L>C_L$ as mentioned in the main manuscript. Interestingly, $T_L>C_L$ seems to be slightly stronger in the mitochondria of Chicken than that of human (or Chimp). We suggest there would be some differences in the mechanism of mtDNA replication across the evolution tree.

Figure 4-figure supplement 2. Correlation of triplet codon frequencies between from observed and from simulated evolutions under the mtDNA somatic mutational signature.

Figure 5-figure supplement 1. Number of recurrent substitutions between silent and missense substitutions. 100 sites were randomly selected from silent substitutions (at 3rd base of triplet codon) and missense substitutions (at 1st and 2nd base of triplet codon). No significant difference was observed among these three groups.

Figure 5-figure supplement 2. Comparison of VAF of protein-truncating mutations (nonsense substitution and indels) across tumor types. Four tumor types with more than 10 protein-truncating mutations are shown. Fisher's exact were applied between breast and other tissue types.

Figure 5-figure supplement 3. Negligible impacts of external mutagens (UV and tobacco smoking) to the somatic mtDNA mutations. No evidence of UV and tobacco smoking was identified even in melanoma and lung cancers, respectively. (Left) We compared the proportion of C>T (and G>A) substitutions in the CpC (GpG) context (mutational signature for UV(Alexandrov et al., 2013)) between melanomas and breast cancers (controls). Because UV

shows trivial impact to the nuclear DNA somatic mutations of breast cancers (Alexandrov et al., 2013), the vast majority of mtDNA C>T substitutions in the CpC context from breast cancers were not generated by UV. (Right) We compared the proportion of C>A (G>T) substitutions between lung and breast (control) cancers. C>A (G>T) substitutions are dominantly generated by tobacco smoking. Like UV, the impact of tobacco smoking to the somatic mutations of breast cancers is trivial (Alexandrov et al., 2013).

Legends of Supplementary Files

Supplementary file 1. Sequencing information of 1,675 tumor-normal pairs.

Supplementary file 2. Catalogues of somatic mutations (substitutions and indels) and inherited polymorphisms identified in this study.

Supplementary file 3. List of phased somatic substitutions.

Supplementary file 4. dN/dS for 13 protein-coding genes in mitochondria.

Supplementary file 5. List of somatic substitution with higher recurrent rate than expected.

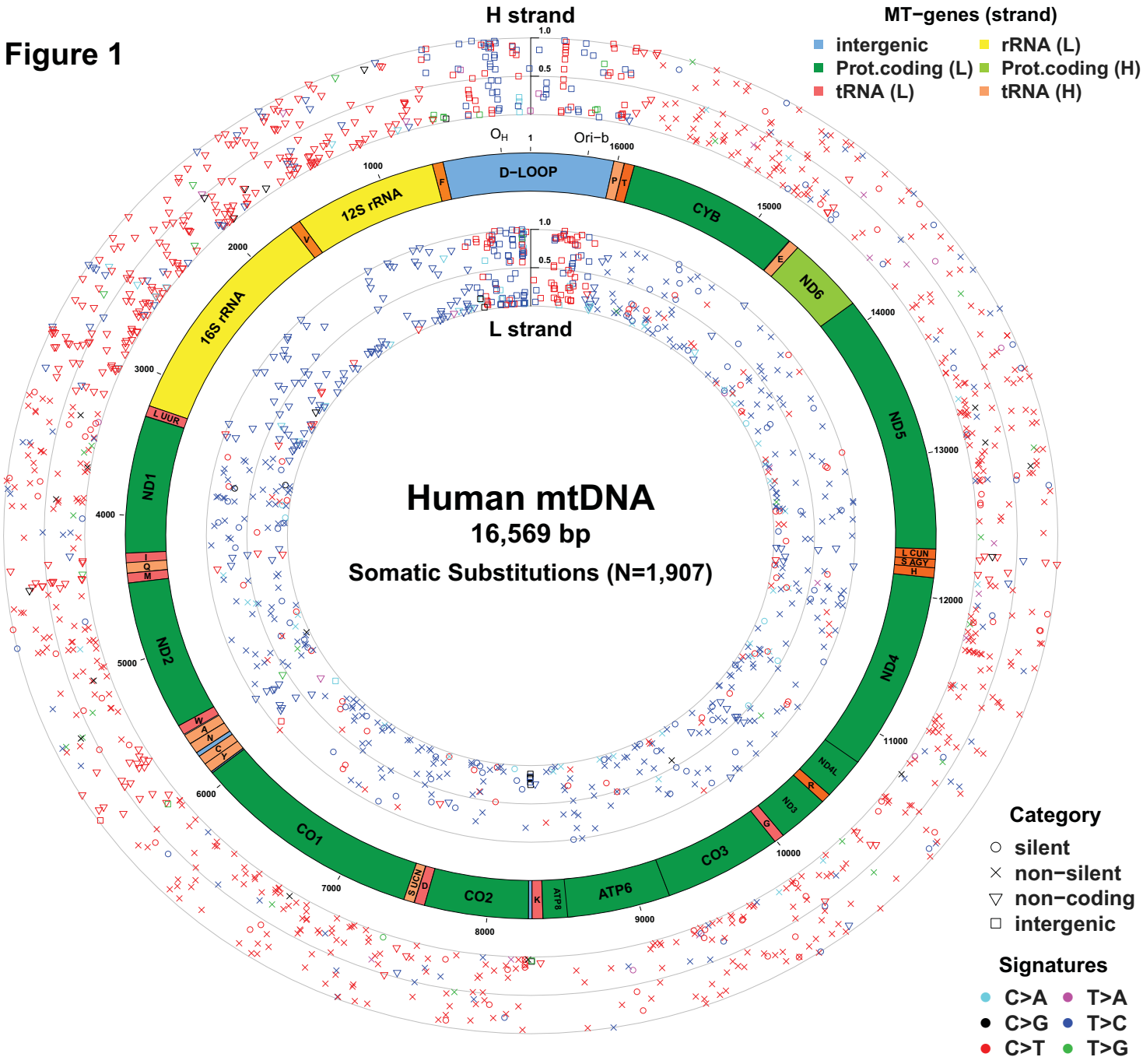
Supplementary file 6. Data accession numbers.

Table 1. Summary statistics of mtDNA sequence data

	WGS	WXS	Average mt RD (WGS)	Average mt RD (WXS)	Total		WGS	WXS	Average mt RD (WGS)	Average mt RD (WXS)	Total
Breast	284	98	11,594.3	52.7	382	Meningioma	0	12	-	42.5	12
Colorectal	1	75	34,916.9	276.6	76	Ependymoma	1	9	10,323.7	52.7	10
Lung	60	0	2,798.1	-	60						
Prostate	80	0	17,810.6	-	80	MPD	12	138	1,517.0	10.9	150
Hepatocellular	0	47	-	205.8	47	MDS	3	75	5,648.7	44.5	78
Melanoma	13	13	513.9	353.5	26	ALL	64	6	886.6	35.9	70
Gastric	0	13	-	184.1	13	CLL	6	0	5,002.2	-	6
Cholangiocarcinoma	0	8	-	143.9	8	AML	1	6	6,783.6	27.4	7
Mesothelioma	0	6	-	106.3	6	Multiple myeloma	0	69	-	43.2	69
Bladder	54	0	646.2	-	54	AMKL	0	9	-	24.2	9
Renal	0	23	-	35.4	23	Lymphoma	0	4	-	99.5	4
Ovarian	0	38	-	58.9	38						
Uterine	27	23	736.0	149.5	50	Osteosarcoma	38	90	9,525.5	119.2	128
Cervical	0	52	-	85.2	52	Chondrosarcoma	0	47	-	99.1	47
Adenoid cystic ca.	1	60	714.7	75.6	61	Ewing sarcoma	0	27	-	69.5	27
Head&Neck	43	3	1,369.1	18.8	46	Kaposi sarcoma	0	9	-	181.0	9
						Chordoma	16	11	1,240.0	82.1	27
Total; 31 cancer types							704	971			1,675

WGS, whole-genome sequencing; WXS, whole-exome sequencing; mt RD, mitochondrial read-depth; MPD, myeloproliferative disease; MDS, myelodysplastic syndrome; ALL, acute lymphoblastic leukaemia; CLL, chronic lymphoblastic leukaemia; AML, acute myeloid leukaemia; AMKL, acute megakaryoblastic leukaemia.

Figure 1


H strand
L strand
MT−genes (strand)
intergenic
Prot.coding (L)
tRNA (L)
rRNA (L)
Prot.coding (H)
tRNA (H)
Human mtDNA
16,569 bp
Somatic Substitutions (N=1,907)
D−LOOP
12S rRNA
16S rRNA
CYB
ND6
ND5
ND4
ND4L
ND3
CO3
ATP6
ATP8
CO2
CO1
ND2
ND1
Ori−b
Category
silent
non−silent
non−coding
intergenic
Signatures
C>A
C>G
C>T
T>A
T>C
T>G

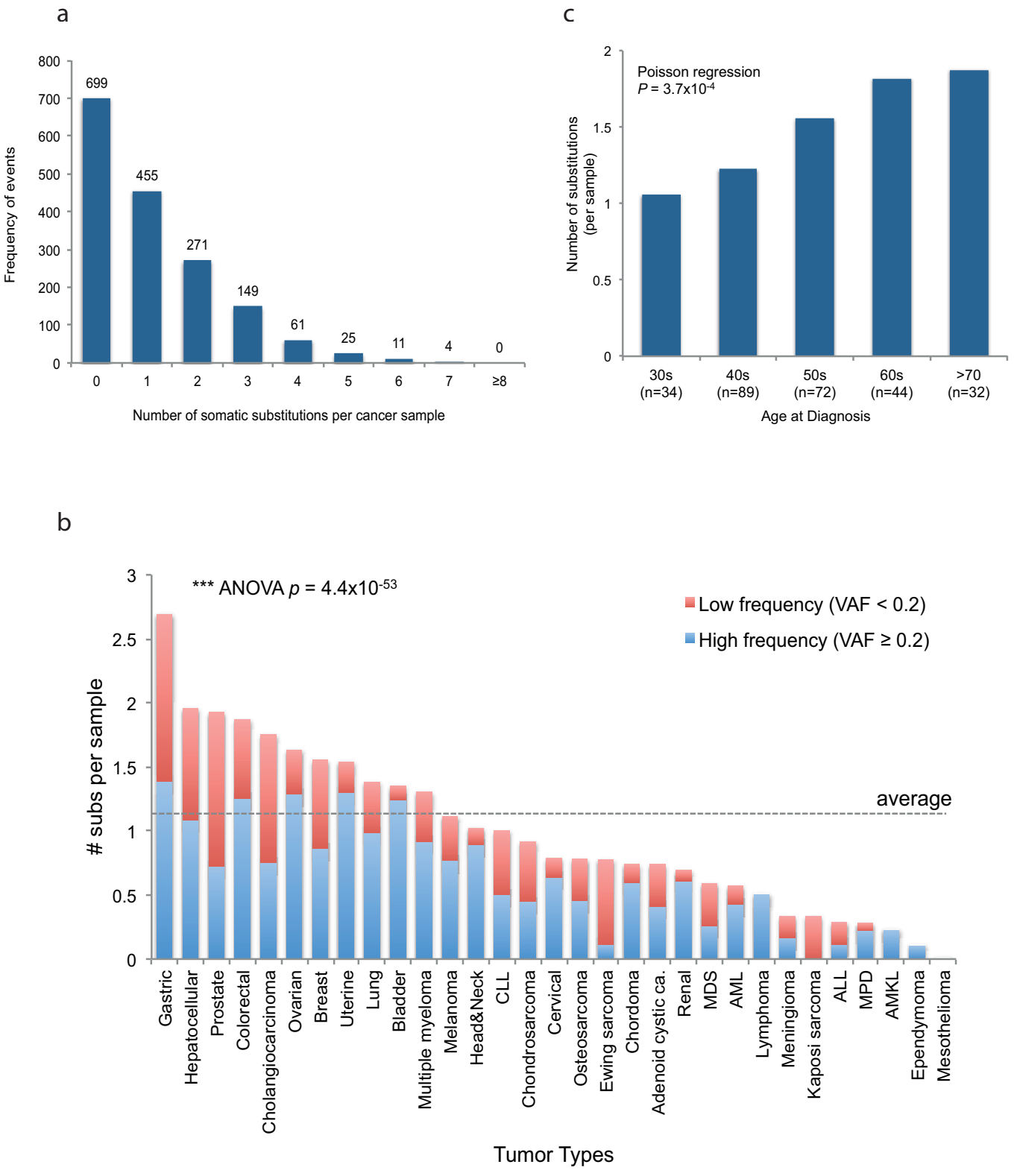
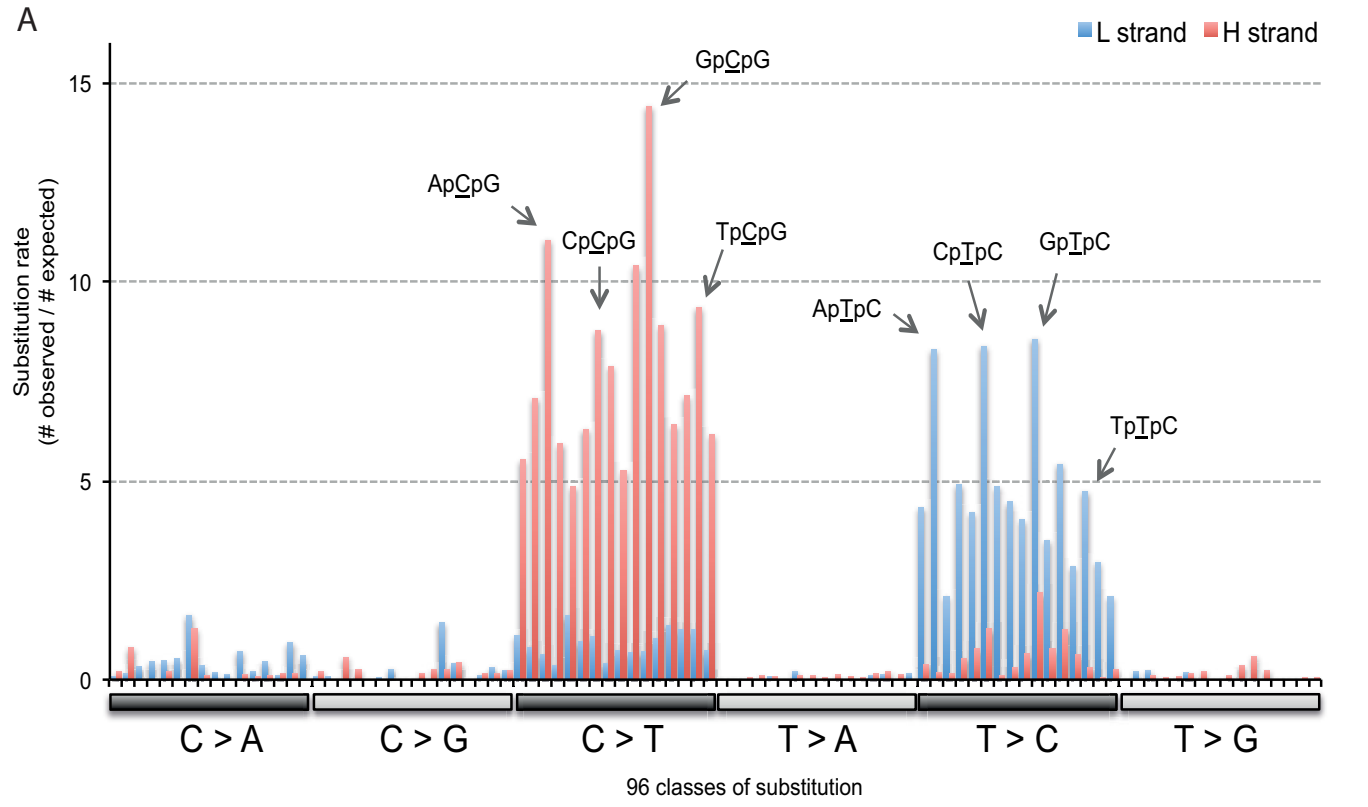
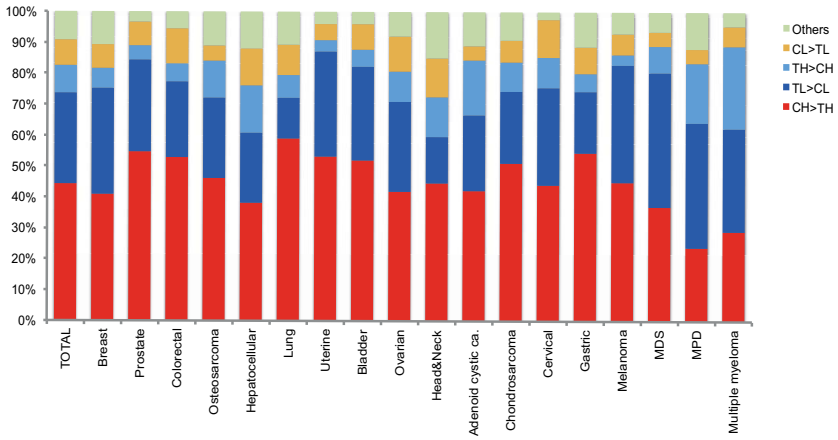

a
Frequency of events
800
700
600
500
400
300
200
100
0
699
455
271
149
61
25
11
4
0
0
1
2
3
4
5
6
7
≥8
Number of somatic substitutions per cancer sample
c
Number of substitutions (per sample)
2
1.5
1
0.5
0
Poisson regression
P = 3.7x10-4
30s (n=34)
40s (n=89)
50s (n=72)
60s (n=44)
>70 (n=32)
Age at Diagnosis
b
*** ANOVA p = 4.4x10-53
Low frequency (VAF < 0.2)
High frequency (VAF ≥ 0.2)
subs per sample
3
2.5
2
1.5
1
0.5
0
average
Gastric
Hepatocellular
Prostate
Colorectal
Cholangiocarcinoma
Ovarian
Breast
Uterine
Lung
Bladder
Multiple myeloma
Melanoma
Head&Neck
CLL
Chondrosarcoma
Cervical
Osteosarcoma
Ewing sarcoma
Chordoma
Adenoid cystic ca.
Renal
MDS
AML
Lymphoma
Meningioma
Kaposi sarcoma
ALL
MPD
AMKL
Ependymoma
Mesothelioma
Tumor Types

Figure 2


A
L strand
H strand
GpCpG
ApCpG
CpCpG
TpCpG
CpTpC
GpTpC
ApTpC
TpTpC
Substitution rate
(# observed / # expected)
15
10
5
0
C > A
C > G
C > T
T > A
T > C
T > G
96 classes of substitution


B
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Others
CL>TL
TH>CH
TL>CL
CH>TH
TOTAL
Breast
Prostate
Colorectal
Osteosarcoma
Hepatocellular
Lung
Uterine
Bladder
Ovarian
Head&Neck
Adenoid cystic ca.
Chondrosarcoma
Cervical
Gastric
Melanoma
MDS
MPD
Multiple myeloma

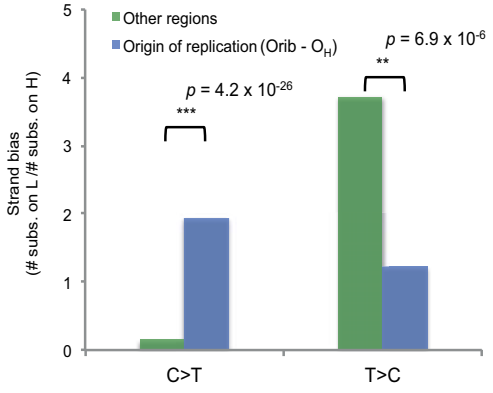
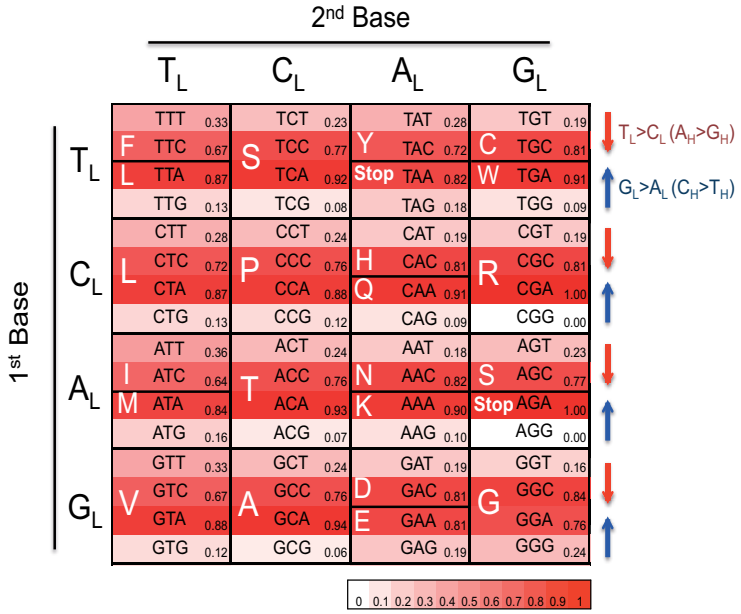

C
Other regions
Origin of replication (Orib - O_H)
p = 4.2 x 10-26
***
p = 6.9 x 10-6
**
Strand bias
(# subs. on L / # subs. on H)
5
4
3
2
1
0
C>T
T>C

Figure 3

A


2nd Base
T_L
C_L
A_L
G_L
1st Base
T_L
C_L
A_L
G_L
TTT 0.33
TTC 0.67
TTA 0.87
TTG 0.13
TCT 0.23
TCC 0.77
TCA 0.92
TCG 0.08
TAT 0.28
TAC 0.72
Stop TAA 0.82
TAG 0.18
TGT 0.19
TGC 0.81
TGA 0.91
TGG 0.09
CTT 0.28
CTC 0.72
CTA 0.87
CTG 0.13
CCT 0.24
CCC 0.76
CCA 0.88
CCG 0.12
CAT 0.19
CAC 0.81
CAA 0.91
CAG 0.09
CGT 0.19
CGC 0.81
CGA 1.00
CGG 0.00
ATT 0.36
ATC 0.64
ATA 0.84
ATG 0.16
ACT 0.24
ACC 0.76
ACA 0.93
ACG 0.07
AAT 0.18
AAC 0.82
AAA 0.90
AAG 0.10
AGT 0.23
AGC 0.77
Stop AGA 1.00
AGG 0.00
GTT 0.33
GTC 0.67
GTA 0.88
GTG 0.12
GCT 0.24
GCC 0.76
GCA 0.94
GCG 0.06
GAT 0.19
GAC 0.81
GAA 0.81
GAG 0.19
GGT 0.16
GGC 0.84
GGA 0.76
GGG 0.24
F
L
L
I
M
V
S
P
T
A
Y
Stop
H
Q
N
K
D
E
C
W
R
S
Stop
G
T_L>C_L (A_H>G_H)
G_L>A_L (C_H>T_H)
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

B

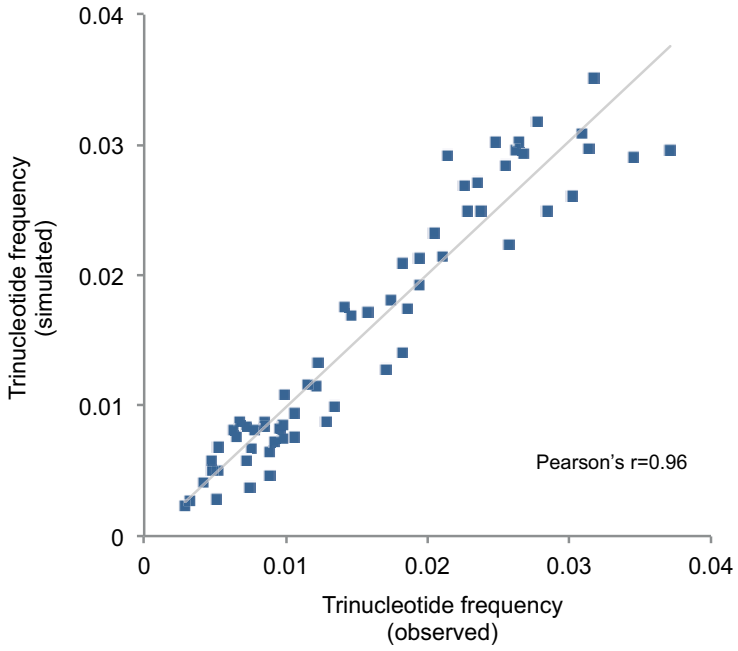

0.04
0.03
0.02
0.01
0
0
0.01
0.02
0.03
0.04
Trinucleotide frequency (simulated)
Trinucleotide frequency (observed)
Pearson's r=0.96

Figure 4

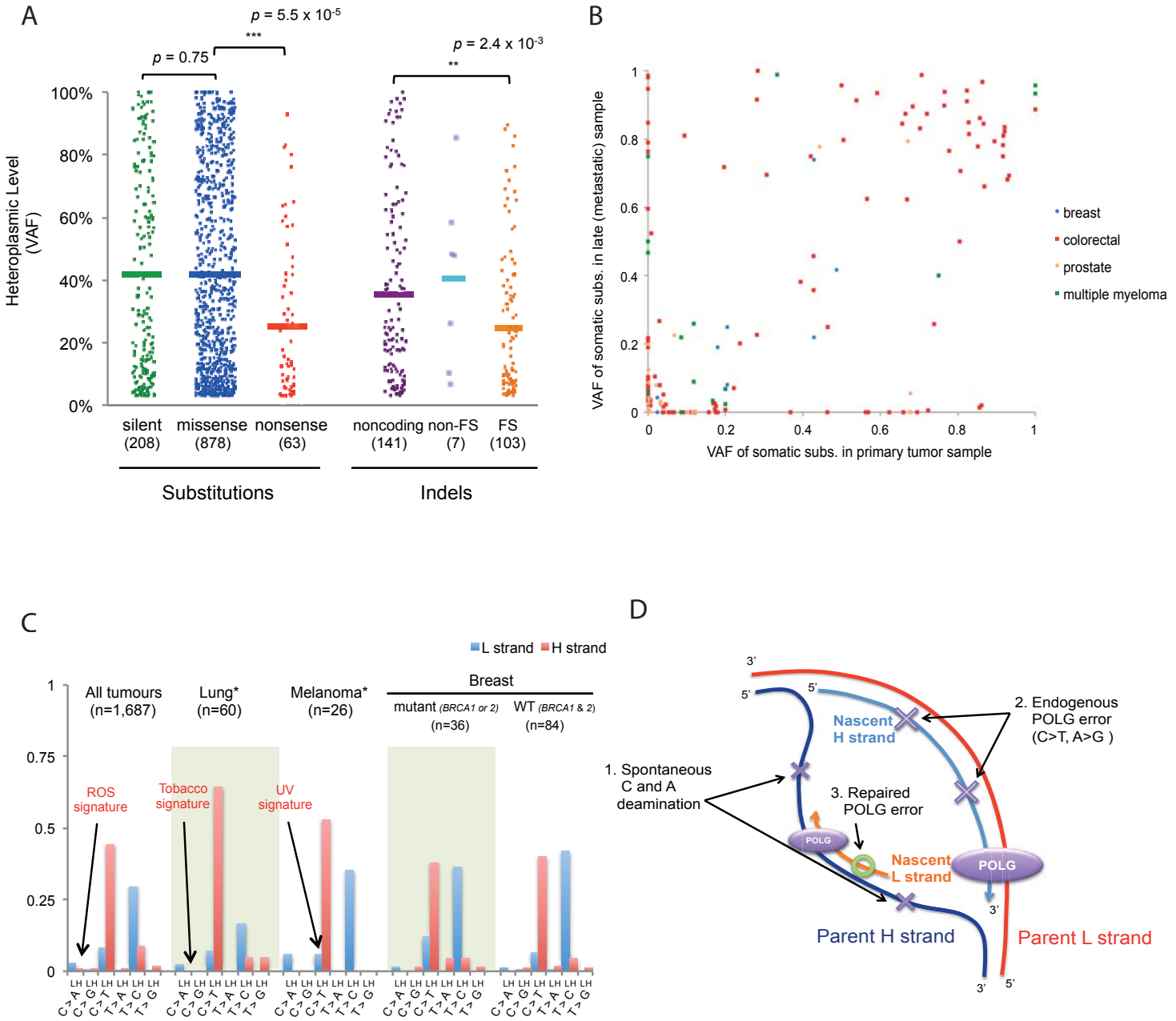

A
p = 5.5 x 10-5
***
p = 0.75
p = 2.4 x 10-3
**
Heteroplasmic Level (VAF)
100%
80%
60%
40%
20%
0%
silent (208)
missense (878)
nonsense (63)
noncoding (141)
non-FS (7)
FS (103)
Substitutions
Indels
B
VAF of somatic subs. in late (metastatic) sample
VAF of somatic subs. in primary tumor sample
breast
colorectal
prostate
multiple myeloma
C
L strand
H strand
All tumours (n=1,687)
Lung* (n=60)
Melanoma* (n=26)
Breast
mutant (BRCA1 or 2) (n=36)
WT (BRCA1 & 2) (n=84)
ROS signature
Tobacco signature
UV signature
D
1. Spontaneous C and A deamination
2. Endogenous POLG error (C>T, A>G)
3. Repaired POLG error
Nascent H strand
Nascent L strand
POLG
Parent H strand
Parent L strand

Figure 5

Participants in ICGC Breast Cancer Group who supplied or verified samples

Elena Provenzano[1], Marc van de Vijver[2], Andrea L Richardson[3,4], Colin Purdie[5], Sarah Pinder[6], Gaetan MacGrogan[7], Anne Vincent-Salomon[8,9], Denis Larsimont[10], Dorthe Grabau[11], Torill Sauer[12], Øystein Garred[12], Anna Ehinger[13], Gert G Van den Eynden[14], C.H.M. van Deurzen[15], Roberto Salgado[29], Jane E Brock[4], Sunil R Lakhani[16,17,18], Dilip D Giri[19], Laurent Arnould[20], Jocelyne Jacquemier[21], Isabelle Treilleux[22], Carlos Caldas[1,23], Suet-Feung Chin[23], Aquila Fatima[3], Alastair M Thompson[24], Alasdair Stenhouse[24], John Foekens[25], John Martens[25], Anieta Sieuwerts[25], Arjen Brinkman[26], Henk Stunnenberg[26], Paul N. Span[27], Fred Sweep[28], Christine Desmedt[29], Christos Sotiriou[29], Gilles Thomas[30], Annegein Broeks[31], Anita Langerod[32], Samuel Aparicio[33], Peter Simpson[18], Laura van 't Veer[34,35], Jórunn Erla Eyfjörd[36], Holmfridur Hilmarsdottir[36], Jon G Jonasson[37,38], Anne-Lise Børresen-Dale[32,39], Ming Ta Michael Lee[40], Bernice Huimin Wong[41], Benita Kiat Tee Tan[42], Gerrit K.J. Hooijer[2]

Affiliations

[1]Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 2QQ, UK

[2]Department of Pathology, Academic Medical Center, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

[3]Department of Cancer Biology, Dana-Farber Cancer Institute, 450 Brookline Ave., Boston, Massachusetts 02215, USA

[4]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, 75 Francis St., Boston, Massachusetts 02115, USA.

[5]East of Scotland Breast Service, Ninewells Hospital, Dundee, United Kingdom

[6]Department of Research Oncology, Guy's Hospital, King's Health Partners AHSC, King's College London School of Medicine, London SE1 9RT, UK

[7]Institut Bergonié, 229 cours de l'Argone, 33076, Bordeaux, France

[8]Institut Curie, Department of Tumor Biology, 26 rue d'Ulm, 75248 Paris cédex 05, France.

[9]Institut Curie, INSERM Unit 830, 26 rue d'Ulm, 75248 Paris cédex 05, France

[10]Department of Pathology, Jules Bordet Institute, Brussels 1000, Belgium

[11]Department of Pathology, Skåne University Hospital, Lund University, SE-221 85 Lund, Sweden

[12]Department of Pathology, Oslo University Hospital Ulleval and University of Oslo, Faculty of Medicine and Institute of Clinical Medicine, Oslo, Norway.

[13]Department of Gynecology & Obstetrics, Department of Clinical Sciences, Lund University, Skåne University Hospital Lund, SE-221 85 Lund, Sweden

[14]Translational Cancer Research Unit, GZA Hospitals St.-Augustinus, Antwerp, Belgium.

[15]Department of Pathology, Erasmus Medical Center, Rotterdam, the Netherlands.

[16]The University of Queensland, School of Medicine, Herston, Brisbane, QLD 4006, Australia

[17]Pathology Queensland: The Royal Brisbane & Women's Hospital, Brisbane, QLD 4029, Australia

[18]The University of Queensland, UQ Centre for Clinical Research, Herston, Brisbane, QLD 4029, Australia

[19]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY, USA
[20]Centre Georges-François Leclerc, 1 rue du Professeur Marion, 21079, Dijon, France
[21]0nstitut Paoli Calmettes, biopathology department, 232 Bd Ste Marguerite, 13009, Marseille, France
[22]Centre Léon Bérard, Lyon, France; Université Claude Bernard Lyon1 - Université de Lyon, Lyon, France.
[23]Department of Oncology, University of Cambridge and Cancer Research UK Cambridge Research Institute, Li Ka Shin Centre, Cambridge CB2 0RE
[24]Dundee Cancer Centre, Ninewells Hospital, Dundee, UK
[25]Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands.
[26]Radboud University, Department of Molecular Biology, Faculty of Science, Nijmegen Centre for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands.
[27]Department of Radiation Oncology, Radboud University Medical Centre, Nijmegen, The Netherlands
[28]Department of Laboratory Medicine, Radboud University Medical Centre, Nijmegen, The Netherlands
[29]Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium
[30]Universite Lyon1, INCa-Synergie, Centre Leon Berard, 28 rue Laennec Lyon Cedex 08 France
[31]Department Experimental Therapy, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands
[32]Department of Genetics, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, O310 Oslo, Norway
[33]Department of Molecular Oncology, BC Cancer Agency, 675 W10th Avenue, Vancouver V5Z 1L3
[34]The Netherlands Cancer Institute, Division of Molecular Carcinogenesis, Amsterdam, The Netherlands.
[35]Department of Surgery, University of California, San Francisco, San Francisco, California, United States of America.
[36]Cancer Research Laboratory, Faculty of Medicine, University of Iceland, Reykjavik, Iceland
[37]Department of Pathology, University Hospital, Reykjavik, Iceland
[38]Icelandic Cancer Registry, Icelandic Cancer Society, Skogarhlid 8, P.O.Box 5420, 125, Reykjavik, Iceland
[39]Institute for Clinical Medicine, Faculty of Medicine, University of Oslo.
[40]National Genotyping Center, Institute of Biomedical Sciences, Academia Sinica, 128 Academia Road, Sec 2, Nankang, Taipei 115, Taiwan, ROC
[41]NCCS-VARI Translational Research Laboratory, National Cancer Centre Singapore, 11 Hospital Drive, 169610, Singapore
[42]Department of General Surgery, Singapore General Hospital, Singapore

Participants in ICGC Chronic Myeloid Disorders Group who supplied or verified samples

Luca Malcovati[1], Sudhir Tauro[2], Jacqueline Boultwood[3], Andrea Pellagatti[3], Michael Groves[2], Alex Sternberg[4,5], Carlo Gambacorti-Passerini[6], Paresh Vyas[4], Eva Hellstrom-Lindberg[7], David Bowen[8], Nicholas CP Cross[9], Anthony R Green[10], Mario Cazzola[1]

Affiliations

[1]Fondazione IRCCS Policlinico San Matteo, University of Pavia, Pavia, Italy
[2]Division of Medial Sciences, University of Dundee, Dundee, UK
[3]Nuffield Department of Clinical Laboratory Sciences, University of Oxford, UK
[4]Weatherall Institute of Molecular Medicine, University of Oxford, UK
[5]Department of Haematology, Great Western Hospital, Swindon, UK
[6]Department of Haematology, University of Milan Bicocca, Milan, Italy
[7]Department of Haematology, Karolinska Institute, Stockholm, Sweden
[8]St James Institute of Oncology, St James Hospital, Leeds, UK
[9]School of Medicine, University of Southampton, Southampton, UK
[10]Department of Haematology, University of Cambridge, Cambridge, UK

Participants in ICGC Prostate Cancer Group

Colin Cooper[1,2,16], Rosalind Eeles[1,3,16], David Wedge[4], Peter Van Loo[4,5], Gunes Gundem[4], Ludmil Alexandrov[4], Barbara Kremeyer[4], Adam Butler[4], Andrew Lynch[6], Sandra Edwards[1], Niedzica Camacho[1], Charlie Massie[7], ZSofia Kote-Jarai[1], Nening Dennis[3], Sue Merson[1], Jorge Zamora[4], Jonathan Kay[7], Cathy Corbishley[8], Sarah Thomas[3], Serena Nik-Zainai[4], Sarah O'Meara[4], Lucy Matthews[1], Jeremy Clark[2], Rachel Hurst[2], Richard Mithen[9], Susanna Cooke[4], Keiran Raine[4], David Jones[4], Andrew Menzies[4], Lucy Stebbings[4], Jon Hinton[4], Jon Teague[4], Stuart McLaren[4], Laura Mudie[4], Claire Hardy[4], Elizabeth Anderson[4], Olivia Joseph[4], Victoria Goody[4], Ben Robinson[4], Mark Maddison[4], Stephen Gamble[4], Christopher Greenman[10], Dan Berney[11], Steven Hazell[3], Naomi Livni[3], Cyril Fisher[3], Christopher Ogden[3], Pardeep Kumar[3], Alan Thompson[3], Christopher Woodhouse[3], David Nicol[3], Erik Mayer[3], Tim Dudderidge[3], Nimish Shah[7], Vincent Gnanapragasam[7], Peter Campbell[4], Andrew Futreal[4,16], Douglas Easton[12,16], Anne Y Warren[13], Christopher Foster[14,16], Michael Stratton[4,16], Hayley Whitaker[7], Ultan McDermott[4,16], Daniel Brewer[1,2], David Neal[7,15,16].

Affiliations

[1]Division of Genetics and Epidemiology, The Institute Of Cancer Research, Sutton, UK

[2]Department of Biological Sciences and School of Medicine, University of East Anglia, Norwich, UK

[3]Royal Marsden NHS Foundation Trust, London and Sutton, UK

[4]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK

[5]Human Genome Laboratory, Department of Human Genetics, VIB and KU Leuven, Leuven, Belgium

[6]Statistics and Computational Biology Laboratory, Cancer Research UK Cambridge Research Institute, Cambridge, UK

[7]Urological Research Laboratory, Cancer Research UK Cambridge Research Institute, Cambridge, UK

[8]Department of Histopathology, St Georges Hospital, London, UK

[9]Institute of Food Research, Norwich Research Park, Norwich, UK

[10]School of Computing Sciences, University of East Anglia, Norwich, UK

[11]Department of Molecular Oncology, Barts Cancer Centre, Barts and the London School of Medicine and Dentistry, London, UK

[12]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK

[13]Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK

[14]Bostwick Laboratories, London, UK

[15]Department of Surgical Oncology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK

[16]Senior Principal Investigators of the Cancer Research UK funded ICGC Prostate Cancer Project

Ethical approval references

Genome Analysis of myeloid and lymphoid malignancies (10/H0306/40)
Genomic Analysis of Mesothelioma (11/EE/0444)
Myeloid and lymphoid cancer genome analysis (07/S1402/90)
The Treatment of Down Syndrome Children with Acute Myeloid Leukemia and Myelodysplastic Syndrome(AAML0431)
CLL (chronic lymphocytic leukaemia) genome analysis (07/Q0104/3)
CGP-Exome sequencing of Down syndrome associated acute myeloid leukemia samples (IRB 13-010133)
Cancer Genome Project - Global approaches to characterising the molecular basis of paediatric ependymoma (05/MRE04/70)
PREDICT-Cohort (09/H0801/96)
ICGC Prostate (Evaluation of biomarkers in urological diseases) (LREC 03/018)
ICGC Prostate (779) (Prostate Complex CRUK Sample Cohort) (MREC/01/4/061)
ICGC Prostate (Tissue collection at radical prostatectomy) (CRE-2011.373)
Somatic molecular genetics of human cancers, melanoma and myeloma (Dana Farber Cancer Institute)(08/H0308/303)
Breast Cancer Genome Analysis for the International Cancer Genome Consortium Working Group (09/H0306/36)
Genome analysis of tumours of the bone (09/H0308/165)

Author Contributions

Y.S.J., M.R.S and P.J.C conceived and designed this study. Y.S.J and M.G performed analysis of the sequence data. Y.S.J and L.B.A analyzed mutational signature. I.M investigated dN/dS ratio of mtDNA substitutions. A.P.B and J.W.T provided bioinformatics support for sequencing data acquisition. M.G. performed the simulation study on evolutionary dynamics. S.N.-Z., M.R., H.R.D., E.P., G.G., A.S., N.B., S.B., P.S.T., J.N., C.E.M., G.S.V., A.R.G., M.-Q.D., A.U., J.E.P., B.T.T., N.M., M.G., P.V., A.K.E.-N., T.S., V.P.C., R.G., J.A.T., D.N.H., D.M., C.F.P., A.Y.W., H.W., D.B., R.E., C.C., D.N., T.V., W.B.I., G.S.B., A.M.F., P.A.F., A.G.L., P.F.C., U.M., contributed samples and scientific advice. Y.S.J, M.G., I.M., M.R.S and P.J.C wrote the manuscript. M.R.S and P.J.C directed the overall research.