Published in final edited form as:
Photochem Photobiol. 2015 January ; 91(1): 15–26. doi:10.1111/php.12377.

UV Signature Mutations †

Douglas E. Brash
Departments of Therapeutic Radiology and Dermatology, Yale School of Medicine, New Haven, CT USA

Douglas E. Brash: douglas.brash@yale.edu

Abstract

Sequencing complete tumor genomes and exomes has sparked the cancer field's interest in mutation signatures for identifying the tumor's carcinogen. This review and meta-analysis discusses signatures and their proper use. We first distinguish between a mutagen's *canonical mutations* – deviations from a random distribution of base changes to create a pattern typical of that mutagen – and the subset of *signature mutations*, which are unique to that mutagen and permit inference backward from mutations to mutagen. To verify UV signature mutations, we assembled literature datasets on cells exposed to UVC, UVB, UVA, or solar simulator light (SSL) and tested canonical UV mutation features as criteria for clustering datasets. A confirmed UV signature was: ≥60% of mutations are C→T at a dipyrimidine site, with ≥5% CC→TT. Other canonical features such as a bias for mutations on the non-transcribed strand or at the 3' pyrimidine had limited application. The most robust classifier combined these features with criteria for the rarity of non-UV canonical mutations. In addition, several signatures proposed for specific UV wavelengths were limited to specific genes or species; non-signature mutations induced by UV may cause melanoma *BRAF* mutations; and the mutagen for sunlight-related skin neoplasms may vary between continents.

Introduction

A mutagen signature, like a person's signature, is unique to its owner. The term “UV signature” seems to have first appeared in a press report (1) on our use of specific UV-induced mutations to deduce the sunlight origin of mutations in human squamous cell carcinoma (SCC) (2). We had been using the word “fingerprint”, but prior to the interview M. Liskay suggested “signature” and the term stuck. The concept itself is much older. UV's predilection for making C→T mutations in viral DNA was shown by Howard and Tessman in 1964 (3), in the same issue of *J. Mol. Biol.* that first described DNA repair replication (4). The simultaneous advent of genetic and chemical methods for DNA sequencing revealed UV's specificity for targeting *E. coli* sites where two pyrimidines (C or T) were adjacent and revealed the unique UV-induced CC→TT substitutions (5, 6). Modifying DNA sequencing

†Presented in part at the 16th International Congress on Photobiology held in Cordoba, Argentina, September 11, 2014, on the occasion of receiving the Finsen Medal of the International Union of Photobiology

Correspondence to: Douglas E. Brash, douglas.brash@yale.edu.

Supplementary Materials: Table S1 can be found at DOI: xxxxxxxxxxx

technology to identify sites of DNA photoproducts revealed a correspondence between mutation hotspots and UV photoproduct hotspots, as opposed to UV just elevating random mutagenesis by inducing the SOS response (7). These experiments all used UVC (100-290 nm). The pyrimidine-pyrimidone (6-4) photoproduct was found to be the main mutagenic lesion in *E. coli*, whereas in human cells and in mouse skin it was the cyclobutane pyrimidine dimer (8-10). Both DNA photoproducts covalently join adjacent pyrimidines, explaining UV's specificity for dipyrimidine sites.

UV signature mutations, defined for the moment as a preponderance of C→T substitutions at dipyrimidine sites, including CC→TT, have been found in the *p53* tumor suppressor gene in SCC and basal cell carcinoma (BCC), actinic keratosis precursors of SCC, and clones of *p53*-mutant keratinocytes in normal sun-exposed skin (2, 11-24). BCC also contain sunlight-induced mutations in *PTCH* (25). SCC from repair-defective xeroderma pigmentosum patients contain a much higher fraction of CC→TT and also contain UV signature mutations in *PTEN* (26-28). Although the role of UV in melanoma was controversial for many years, melanoma cell lines contained UV-like mutations in the cell cycle regulator *CKDN2* (29). Next-generation sequencing of melanomas from sun-exposed body sites has now revealed UV signatures in many genes, including the apparent oncogene *RAC1* and the apparent tumor suppressor *PPP6C*, with an overall hotspot motif very much like the 5' $(A:T)_nTC$ 3' motif seen in *E. coli* (7, 30, 31).

Excellent surveys describe the mutations seen in various human tumors and their presumptive mutagens (32, 33). However, identifying a set of “tumor signatures” does not itself identify the mutagen. Interpreting tumor mutations requires inference from studies of mutations caused by known mutagens. This review focuses on ultraviolet radiation, for which the UV sources are UVC, UVB (290-320 nm), UVA (320-400 nm), or solar simulator light (SSL, ∼ 310-1000 nm; typical proportions are 0.8% UVB, 6% UVA, 43% visible light and 47% infrared. The proportions for sunlight at the earth's surface are ∼ 0.3%, 5.1%, 62.7% and 31.9% (34)).

Canonical Mutations *vs* Signature Mutations

It's one task to sequence mutations in cells treated with a known mutagen, and conceptually quite another task to deduce the mutagen from the tumor mutations. This is simply because mutations typical of one carcinogen can also result from another carcinogen (Fig. 1). Before analyzing these UV signature mutations, we first clear up several misconceptions common in the cancer literature and outline how mutation signature mutations are properly employed.

The term *mutation* is misused surprisingly often in cancer biology. A chemical change to the DNA is not a mutation but is *DNA damage* or a *DNA lesion*, or more specifically a *DNA adduct* (when extra atoms have been added to the DNA) or a *DNA photoproduct* (when UV has rearranged the bonds between existing atoms). The DNA at a damage site is no longer truly DNA. Once the DNA strand is copied, during repair or during S phase DNA replication, the daughter strand is once again normal DNA but the sequence may be incorrect; this base change is the mutation. A single base change is a *base substitution*, a type of *point mutation*; single-base or longer insertions or deletions can occur, and there can

be gene rearrangements. Mutations also arise in mRNA during transcription past a DNA damage site (35), but these lie outside our scope. Typically a lab measures *mutation frequency*, the fraction of new mutants in a population of cells. Depending on the assay, this value is expressed as mutations per base or mutations per gene. (Expressing it per base or gene suggests that measuring the frequency of mutations at one base in many cells is equivalent to measuring the frequency across many bases in one cell, which is only true on average.) Because mutant cells create mutant daughters, it is important not to overestimate this number by including the progeny in the mutation frequency. In the lab, progeny are ignored by seeding cells at a density low enough that each founder mutant and its progeny give rise to a single colony. This genetic approach also solves the problem that mutation frequencies are very small; after UV, any particular base is mutated in one cell in 10^6 or less. In blood, skin, or a tumor we don't know which mutant cells were founders and which were daughters (although see (36)); we will term this combined number *mutation prevalence*. Some laboratories measure the more difficult *mutation rate*, the number of mutations per base per cell division. A final point is that mutagens are usually also toxic, so mutation numbers are actually mutations per survivor. Whether surviving cells differ from those that died is an unresolved question, although it will eventually be answered by deep sequencing for unselected rare mutations without first removing the vast excess of normal cells (37, 38).

Sequencing the mutations within a gene reveals their spatial distribution across the gene, typically showing that they cluster at hotspot bases having a mutation frequency 3-10 fold higher than elsewhere. Sequencing also reveals the kinds and relative proportion of mutation types, such as C→T or T→G (more precisely, G:C→A:T etc., unless we know which DNA strand contained the damage). Both the spatial and mutation type distributions have been given the name *mutation spectrum*, so we will introduce the phrase *mutation log* to simply mean all of the mutations observed after treating cells with a known mutagen, including derived characteristics such as spatial distribution, mutation types, predilection for flanking bases or 5-methylcytosine, or any other features.

To deduce the mutagen from a dataset of tumor mutations, we need to use a subset of the mutagen's mutation log that is caused *only* by that mutagen. This subset is the *mutation signature*. The rest of the mutagen's mutation log is not so distinctive; these will be termed *non-informative* in the technical sense used in human genetics: it is accurate information, and these mutations were caused by the mutagen, but they cannot be used for inference purposes. A mutagen therefore produces *signature mutations* plus *non-informative mutations*. Mutations in the tumor's mutation log that resemble a mutagen's non-informative mutations – or which resemble a mutagen that has no signature – can be consistent with that mutagen, but they are not conclusive.

The mutation signature is therefore a computed concept, ascertained by comparing with mutations from other mutagens. Yet our accumulated store of “distinctive” mutation patterns usually comes from papers that study a single mutagen and notice that its mutations depart in a particular way from a random distribution of base changes. This pattern will be termed the mutagen's *canonical mutations*. A single-mutagen view is excellent for asking questions about that mutagen, such as how its DNA lesions are processed by a DNA polymerase that makes mistakes. But a canonical mutation is usually not unique to that mutagen. The

mutation signature is therefore constructed from a subset of the canonical mutations and is validated by comparing it to mutations made by other mutagens.

A moment's reflection reveals that mutation signatures have important properties that must be taken into account when they are used:

- A mutation signature is really a matter of degree, since no mutation is completely unique. Most mutation types can be caused by a variety of mutagens, so it is almost never possible to ascribe a cause to a single occurrence of a mutation. A CC→TT substitution after UV is perhaps the closest to a perfect signature, but even it can be found in very low amounts after treating with other mutagens if the detection assay has been designed to detect only CC→TT changes (39, 40).
- Many mutagens, such as oxidative agents, have canonical mutations that distinguish them from unrelated mutagens such as UV or polycyclic aromatic hydrocarbons, but do not distinguish them from each other. Ultimately, a mutation signature is a probabilistic concept related to information theory in that it depends on knowing the mutation logs of all the other possible mutagens. We do not pursue this avenue here.
- Canonical mutations are usually defined by an atypical frequency of a mutation type in a *pool* of mutations measured over many cells, many genes, or tumors from many patients. An example is the frequent C→T substitution at a dipyrimidine site after UV. But it is essential to realize that any C→T has a 75% probability of being part of a dipyrimidine just by chance (1 – (50% chance of a purine on its 5' side * 50% on its 3' side)), so “atypical” means significantly greater than 75%. Moreover, the backward inference using a mutation signature also applies to a pool. Therefore the mutation present in any single tumor is not conclusive. This fact has legal ramifications for occupational exposure to carcinogens.
- Only some of the mutagen's mutation log will be mutation signature mutations. Canonical mutations are not always frequent. CC→TT changes are present at 5-15% of mutations and are sometimes not seen at all. The higher the fraction, the more useful the signature. The CC→TT substitutions by themselves leave open the possibility that the other mutations in the dataset came from another mutagen. Most of the UV signature's work is done by the less specific but more prevalent C→T substitutions and their dipyrimidine location.
- When mutation signature mutations are seen in a collection of tumors, the carcinogen's noninformative mutations must be there as well! The presence of non-signature mutations does *not* imply that a second mutagen was acting: a second mutagen becomes likely only if we observe unexpected types of mutations or a higher-than-expected fraction of non-signature mutations. For example, SCCs contain, in addition to their UV signature mutations in *p53*, a substantial proportion of oxidative-looking mutations (2). But these are expected: they are the non-informative mutations seen in UV-irradiated cells in culture. Whether they originate from UV-induced oxidative damage or from odd behaviors of an error-prone polymerase acting at cyclobutane dimers is unknown. BCCs contain an even

higher proportion of oxidative-looking mutations (25); the ones in excess of those expected from the non-informative mutations are a clue that there is a second mutagen. But in deducing the initiating mutagen for a tumor or mutated tissue, it is not legitimate to ascribe only the signature mutations to their mutagen and ascribe the remaining mutations to a different mutagen (41).

- A mutation signature can be overturned tomorrow if a new mutagen is discovered that creates the same type of mutations.

A third caveat in employing mutation signatures – in addition to non-informative mutations and pools – is phenotypic selection. When selection is imposed on the mutagen's handiwork, as happens with a lab's drug selection or in a tumor, the base substitutions are viewed through two veils. The first veil is that some base substitutions will not change the amino acid, due to the genetic code. The second veil is that only some amino acid changes can change the protein's phenotype. For a gene that is scored by its loss of function, such as *lacI*, *HPRT*, or a tumor suppressor gene, many changes are effective: stop codons, amino acid changes (missense mutations), and splice site mutations. Moreover, different base substitutions in a codon often lead to an amino acid change and many sites in the gene can prevent the protein's function. But for a gene scored by gaining a new function, such as the *RAS* oncogene, only certain amino acids will provide the new function and they can be reached only by specific base changes. In the extreme case, phenotypic selection choses the carcinogen's non-informative mutations because those are the only ones that will achieve the needed amino acid. For gain-of-function mutations, it becomes dubious to deduce the carcinogen from the tumor's mutation. We just cannot know.

A case in point is the *BRAF* oncogene, mutated in melanomas and benign nevi. The common Val600Glu T→A mutation is not a UV-like base change, nor does it occur at a dipyrimidine site, so for a decade the field concluded that UV was not an important mutagen for melanoma. Whole-exome and whole-genome sequencing of melanomas have now revealed a preponderance of UV signatures in melanomas from sun-exposed body sites (30, 31), reminding us that even the *BRAF* mutations could have been caused by non-informative UVinduced mutations – notably the 10% of mutations seen in UV mutagenesis experiments that are T→A, some of which are not at dipyrimidine sites. This possibility is underscored by *BRAF* mutations that are part of a tandem double mutation, consisting of the Val600 T→A plus an adjacent base change that does occur at a dipyrimidine site (42). Tandem mutations besides CC→TT are often seen in UV mutation studies, albeit at low frequency, and are considered evidence of polymerase errors extending beyond the photoproduct site.

Mutation signatures are a blunt instrument because most mutagens create mutations similar to those from other mutagens. However, the tool is sharper for some mutagens including psoralen (43), aflatoxin B_1 plus HBV virus (44), and UV. Having discussed the logical pitfalls in finding and using mutation signatures, we now proceed to our meta-analysis of the existing UV mutation data.

Signatures of UVC, UVB, UVA, and SSL

Several excellent reviews discuss UV signatures, often in connection with proposing a new one (45-49). For the present meta-analysis, we focus on studies meeting three criteria: i) mammalian systems, including shuttle vectors, endogenous genes in cultured cells, or transgenes in mouse skin; ii) systems capable of registering all base changes (notably excluding assays limited to nonsense mutations (50, 51)); and iii) experiments in which more than 12 mutations were accumulated, resulting in some pioneering papers being omitted (52-55). Included in the analysis were 17 datasets with UVC exposure (56-72), 10 with UVB (60, 73, 68, 74-80), 11 with UVA (73, 68, 74, 81-83, 80, 84-87), and 6 with SSL or sunlight (34, 68-70, 88), as well as 4 from sun-related skin lesions (2, 14, 15, 41) and, as controls, 1 H_2O_2, 1 benzo(a)pyrene, and 5 aflatoxin B_1 (89-94). For each paper and wavelength, we computed mutation features such as the base change and sequence context. Data were converted to fraction of total mutations in order to correct for different experiment sizes.

Target-gene differences

It rapidly became clear that a commonly used metric for UV-induced mutations, the frequency of mutations that could have occurred at the 3' base of the dipyrimidine, varies widely with the gene being assayed. This is due to variations in the prevalence of pyrimidine runs, whose internal sites will be non-informative with regard to the 5' versus 3' location of the mutated pyrimidine. Prokaryotic gene targets, even as mammalian transgenes, were often mutated at isolated dipyrimidine sites or in short runs, but eukaryotic genes tended to be mutated within pyrimidine runs. Correspondingly, the most frequent C-containing cyclobutane pyrimidine dimer sites in six mammalian genes occurred in long pyrimidine runs (95). Therefore, a bias for 5'PyC3' sites supports a UV origin, but absence of a 3' bias may just mean that the genes examined have many pyrimidine runs (making the gene non-informative in this regard).

Similarly, the frequency of available CpG targets varies widely between genes, with *lacI*, *aprt*, *hprt*, *Hprt*, and *HPRT* rarely showing mutations at PyCG sites. It is difficult to know whether this variation is due to the number of available CpG sites, as one must also know which CG sites can cause a change in amino acid sequence that can also be scored as a change in protein phenotype; these details have been gathered for *p53* (87). Therefore, for our analyses, the category proposed in the literature, “C→T mutations at PyCG sites”, was modified to “mutations at CG that are PyCG→PyTG”.

We therefore focused on 9 additional mutation properties that have been proposed to be diagnostic: location at a dipyrimidine site (diPy), transition mutation (Py→Py), C→T, C→T that are at diPy, C→T & at diPy, C→A, T→G, fraction of mutations at CpG that are PyCG→PyTG, and mutant Py located on the non-transcribed strand (NTS). For brevity, the mutations are expressed as if they arose at the pyrimidine because the UV DNA photoproduct is on the pyrimidine-containing strand. H_2O_2, benzo(a)pyrene, and aflatoxin B_1 are known to react with the purine, but abbreviating the mutation as above does not affect the mutation categories. Two classes of mutations primarily seen with UV were excluded,

due to their rarity and diversity: tandem double mutations other than CC→TT and non-tandem double mutations, also termed “triplets” (57, 47).

Canonical mutation patterns

The mutation fraction data for all the datasets and mutation categories can be compared by visualizing the data as a heat map, an approach common in genomic analyses. Red represents a high percentage of the property, blue represents a low percentage, and black indicates missing values (Fig. 2; see Methods). Several patterns are apparent by inspection.

- UVC, UVB, UVA, SSL (a mixture of UVB and UVA), and skin lesions from sunlightexposed body sites are broadly similar, with a high percentage of transition mutations, C→T substitutions, C→T that are at diPy, and C→T & at diPy, of mutations at dipyrimidine sites, and of mutations at CG that are PyCG→PyTG. The percentage of C→A or T→G is generally low.
- The inverse pattern is seen for benzo(a)pyrene, aflatoxin, and H_2O_2, so the above-mentioned criteria have the power to discriminate UV from these three chemical mutagens.
- The photobiologist's favored criterion “C→T that are at diPy” has the highest positive signal. It also has the attractive property that any value above 0.75 is non-random (2). However, it has the weakness that the rare C→T produced by AFB_1 do often occur at diPy, giving this metric a value similar to that of UV. Combining the two canonical patterns as “C→T & at diPy” gives a lower frequency but one that consistently classifies UV and the chemicals into different groups. This canonical pattern is therefore a useful UV signature.
- Three exceptions to these rules are remarkable. First, the wild type hamster cell lines V79 and AA8 give a pattern of UVC mutations that resembles the H_2O_2 and AFB_1 patterns, although their repair-defective analogs do not. This fact hints that hamster cells make atypical mutations during mis-repair of guanine. UVA-exposed CHO hamster cell lines were similar. Second, UVA-induced mutations in some experiments (83, 84), which will be termed “UVA group II”, resemble the mutations produced by H_2O_2, UVA+riboflavin, or UVA+porphyrin; they have a high percentage of G→T transversions (C→A) and these are not at dipyrimidines. This behavior has been attributed to irradiating the cells in medium, which contains photosensitizers such as riboflavin (49). Third, skin lesions from Australia resembled the oxidative pattern from H_2O_2, UVA+riboflavin, and UVA+porphyrin, rather than the UV-like pattern of sun-related skin lesions from American and Swedish patients. This fact hints at differences in photochemistry at different latitudes or UV doses, or the presence of different photosensitizers in different skin types.
- A property commonly cited for UV-induced mutations is a bias for the pyrimidines being on the non-transcribed strand (30, 33), due to transcription-coupled repair of the cyclobutane pyrimidine dimer. But this strand bias is not a general property of UV-induced mutations. The heat map makes it clear that repair-defective cells do not show a strand bias for mutations. Nor do the *supF*, *cII*, *gpt*, or most *lacZ*

transgenes in mammalian cells, even repair-proficient ones. This behavior is reasonable if the viral and bacterial genes are silent in the mammalian cell and are only expressed in the subsequent bacterial assays used to detect the mutations. The same set of genes displaying strand bias holds for H_2O_2, benzo(a)pyrene, and AFB_1, but the DNA lesion from these compounds is on the purine-containing strand. In a tumor, UV mutations are expected to exhibit a strand bias because the tumor's oncogenes and tumor suppressor genes are presumably transcribed. But in whole-exome or whole-genome sequencing, the average strand bias in the 2% of expressed genes can be overwhelmed by the majority of non-expressed genes (30). In sum, the presence of strand bias supports a UV origin, but a) it is not limited to UV and b) absence of strand bias in a large collection of genes is non-informative.

- T→G substitutions have been proposed as a signature of UVA and SSL (73). This pattern clearly holds for wild type CHO, V79, and AA8 hamster cells, but it is not prevalent in other species or even in repair-defective hamster cells. Using T→G substitutions in the *p53* gene in skin lesions as a way to assign their origin to UVA rather than UVB (41) therefore lacks a firm basis; a stronger inference for this Australian skin lesion dataset is that several of the mutation types resemble those of oxidative or photosensitizer origin.
- C→T substitutions are elevated at PyCG, partly due to the enhanced cyclobutane dimer formation at methylated cytosines, especially for UVA (81). The "C→T at PyCG" pattern might be proposed as a signature of UVA, but its frequency with UVA is 83% versus 58% for UVB. It is also apparent from the heat map that, in genes with an adequate number of CpG sites such as *supF*, "C→T at PyCG" mutations are seen even with UVC. Therefore the presence of "C→T at PyCG" mutations seems to be a canonical property of ultraviolet light, instructive about the origins of UV mutations at CG sites, but it would require very fine tuning to be used as a signature for deducing from the mutations the specific UV wavelength that caused a tumor.

Clustering analysis

We next asked whether unbiased clustering techniques identify distinct signatures for different UV wavelengths, or at least numerical differences in proportion between wavelengths (see Methods). First, we utilized hierarchical clustering because this algorithm does not require imputation of missing values. In a small data set such as this one – ∼50 datasets each containing ∼40 data points – only the first few levels of clustering will be meaningful. Independent of the distance measure used, we found (not shown) that: the AFB_1 and BPDE data clustered together; the H_2O_2 data and the UVA data from (82-85), including group II, clustered together and lay in the same meta-cluster as the AFB_1 and BPDE data; and the UVC, UVB, UVA, and SSL data clustered together, except for hamster data sometimes clustering with H_2O_2. K-means clustering allows robustness to be assessed by repeatedly clustering the same data using different random-number seeds. However, for this algorithm it is necessary to impute missing values such as for the criterion "mutations at CG that are PyCG→PyTG" (at genes with few mutable CpG sites). For imputation, we used the average value for the dataset at CpG-containing genes, 0.75. Different seeds produced

different clusters but, whether the number of clusters stipulated was 2, 6, or 12, the same clustering tendencies were seen as for hierarchical clustering (not shown).

Non-negative matrix factoring (NMF) consensus clustering has the distinct advantage that the robustness of clusters is assessed by quantifying their reproducibility over multiple iterations. In a graphical plot of NMF clusters (Fig. 3), a dataset that sometimes falls into one cluster and sometimes into another appears at the edge between the two groups. The degree of dispersion between iterations is quantified as the cophenetic correlation coefficient; a maximum of ∼ 0.99 was typically reached when k, the stipulated number of clusters, was 3. When k was 2 or 3, respectively, two or three stable groups were readily identifiable (dark red). These clusters were (bottom to top): a) repair defective cells (UV) + UVC + UVB + SSL + Sun (US/Sweden) + UVA group I; b) UVA group II + UVA-riboflavin + Sun (Australia) + UVB/SSL hamster + BPDE + AFB_1; and c) UVA group II + UVA-porphyrin + H_2O_2 + Sun (Australia) + UVC/UVA hamster. The clusters were separated from each other by off-diagonal bands representing datasets that clustered into one or the other of its neighbors on different runs (green, orange, and light blue bands). These unstable datasets included hamster data and Australian normal skin mutations. The last two groups of data sets often clustered with each other in different runs, but never with the first group (dark blue). Beyond k = 2-3, NMF clustering fragmented natural groups into subsets that had no obvious commonality and varied between runs, resulting in low cophenetic scores.

We then asked whether a subset of mutation features – the putative “UV signature” – suffices to distinguish UV from non-UV mutagens. Using the two canonical UV features of “dipyrimidine site” and “C→T” (Fig. 4a) yielded clusters that showed much greater dispersion and it intermixed datasets that had been separated in the NMF analysis when using 9 features (Fig. 3), although UV datasets were still separate from BPDE and AFB_1 datasets. Adding two additional UV-like features, “transition mutations” and “CG mutations that are at PyCG>T”, improved the cophenetic score but substantial dispersion remained (Fig. 4b). However, strikingly better resolution – even revealing 3 clusters – was obtained by instead supplementing the two original canonical UV features with two negative controls: two features rare with UV (A:T→C:G and G:C→T:A), the latter of which is a common chemical mutagen signature (Fig. 4c).

Quantification

The UV signature can be refined by adding numerical values from the literature (Table S1; see Supporting Materials): ≥60% of all mutations are C→T at a dipyrimidine site, with ≥5% being CC→TT. The median of “C→T & diPy” ranged from 62-75% of total mutations, with hamster cells at the low end; notable exceptions were ∼25% for UVA group II and Australian skin lesions. CC→TT were 5-10% of total mutations, with large fluctuations; the high end came from repair-deficient cells and the lowest from UVA, *lacZ*, hamsters, UVA group II, and Australian skin lesions.

Other canonical UV features were less distinctive. The premier example is that ≥98% of C→T mutations after UV occurred at dipyrimidines, far in excess of the well-defined 75% threshold expected by chance. But this criterion was less distinctive because the percentage

was also high for AFB_1, which made few C→T but these were at dipyrimidines (Fig. 2). "C→T" alone (without the diPy restriction) had a median frequency of 62-83% of total mutations in different groups, with the low and high ends again from hamsters and repairdeficient cells, respectively, and the exceptions again from UVA group II and Australian skin lesions at ∼33%. The "diPy" feature behaved similarly. The median of total mutations that were diPy ranged from 92-99% and the median of "C→T that were diPy" ranged from 98-100%. The usual groups were exceptions, with UVA group II and Australian skin lesions being at or near the random level of 75%. For the methylation-based criterion "percentage of C→T that are at CG", the median was 24-39% for UVC, UVB, and SSL (excluding the *lacI*, *aprt*, and *hprt* genes, which do not report CpG mutations well) and doubled to 61% with UVA, consistent with the CpG preference being a canonical UVA feature. It was only 44% in UVA group II and was 0% in Australian skin lesions.

Differential diagnosis is possible because the proportions are quite different from chemicals. BPDE and AFB_1 were ∼60% G→T, with G→C also common and no bias toward dipurines. UVA plus the photosensitizer riboflavin also gave ∼60% G→T. UVA plus porphyrin, on the other hand, led to ∼40% G→A (and C→T, with no preference for diPy), plus T→G and other transversions. H_2O_2 was intermediate between UV and chemicals, with 50% of total mutations being C→T at a diPy site; but there was no preference for diPy sites and the next most frequent base change was T→G. UV-irradiated hamster cells often resembled the H_2O_2 or AFB_1 patterns.

Discussion

These analyses lead to conclusions, lessons, and caveats. The overarching conclusions are: a) UVC, UVB, UVA, and SSL mutations are generally similar. b) Numerical differences in the proportion at CpG constitute a canonical pattern for UVA, but only certain genes are good reporters. Hence using this feature to reliably deduce the mutagenic wavelength from a tumor's mutations would require careful validation of the genes, while distinguishing between 30% and 60% incidence would require large sample sizes. c) The canonical UV mutations are distinct from those created by BPDE, AFB_1, H_2O_2, or UVA irradiation of photosensitizers. d) The canonical UV pattern "≥ 60% C→T substitutions at a dipyrimidine site" constitutes a UV signature, as does "≥5% CC→TT". e) Less distinctive canonical UV mutation features were: ≥98% of C→T occurring at dipyrimidines, which also held true for a chemical mutagen making few C→T; a bias for the mutated pyrimidine lying on the non-transcribed strand, which tends to apply to active genes and applies to non-UV mutagens; a bias for C→T substitution at CpG sites, which applies best to methylated genes and these tend to not be expressed; and a bias for mutation at the 3' pyrimidine, which is obscured in eukaryotic genes, which tend to contain runs of adjacent pyrimidines. Individually, each feature can distinguish between many UV- and non-UV mutation datasets. But using the UV signature to ask whether a collection of mutations can be classified as having originated from UV is markedly sharpened by also testing for the *non*-UV canonical mutation types that should be rare or absent. The last finding recalls Aristotle's dictum that a definition should state what a thing is and what it is not.

A salient lesson is that each gene targeted is different. First, prokaryotic genes differ from mammalian genes in not having mutations targeted to splice sites or CpG. This was a stroke of luck in our initial experiment correlating DNA photoproducts with mutation frequency (7), since mutations dependent on cytosine methylation might have obscured the linear relationship. Even as a transgene in mammalian cells, *lacI* rarely shows mutations at its PyCG sites, but lambda *cII*, *E. coli supF*, and some *lacZ* gene constructs are methylated in mammalian cells and report CpG sites quite well. Conversely, the endogenous mammalian genes *aprt*, *hprt*, *Hprt*, and *HPRT* rarely showed mutations at PyCG sites; most of their CpG sites are in unmethylated CpG islands and few of these would alter the amino acid after a C→T change. *p53* is methylated and its mutation hotspots report C→T mutations at CpG. The physical basis for the CpG preference involves a 3-15 fold greater initial cyclobutane dimer formation at methylated C, especially for UVB and SSL (96, 97) and the million-fold faster deamination of C or 5MeC in cyclobutane dimers to a mutagenic U or T (98-103).

Another bacterial property is that the mutations often occurred at isolated diPy or at the 3' end of a pyrimidine run, whereas in mammalian genes they often appeared inside a run. As a result, the oft-cited ratio of 3'/5' locations for the mutated pyrimidine was 3-50 fold in bacterial genes, even when present as transgenes in mammalian cells, whereas the ratio in mammalian *aprt* or *HPRT* was often 1 or less.

Future mutation signature experiments should be guided by several caveats. First, it is essential to report the buffer present during UV irradiation. This should be a saline solution because irradiating medium components such as phenol red, riboflavin, and tryptophan triggers photosensitization reactions that, it is now clear, entirely alter the mutation log (Fig. 2 and (82-85)). Another caveat is that the shortcut of sequencing a mutant clone's mRNA, rather than the inconveniently scattered exons of genomic DNA, will tend to miss stop codon mutations because these often lead to nonsense-mediated mRNA decay (104). This shortcut will also miss UV-mutated splice sites and branch sites which, when searched for in *HPRT*, were found often (67, 66). Finally, any particular gene's canonical patterns are dominated by the gene's hotspots. For example, the *supF* CpG site mutations are due to its hotspots at nucleotides 155 and 156, and over half of those in *cII* are at the nucleotide 196 hotspot. If the mutation hotspots reflect damage hotspots, then the clustering of mutations into hotspots does reflect the molecular mechanism. But if hotspots reflect phenotypic selection, we need to be cautious. The same over-representation of hotspots will be true of any of the generalizations we have made above, so comparing several genes is always wise.

The search for signatures that discriminate one mutagen from another can be improved further. Clustering methods merely explore the structure of the datasets in the space defined by the features that were chosen for examination. The next step would be discriminant methods, such as Principle Components Analysis, which would reveal the effectiveness of each combination of canonical features in separating groups of data arising from different mutagens. These analyses will require datasets from the non-UV mutagens that are as extensive as the ones assembled here for UV.

A direction for the future is to include phenotypic selection explicitly, by a rigorous analysis that divides M_i, the mutation frequency at each nucleotide i, into the product of preselection

mutagenesis P_i × the selection coefficient W_i for each nucleotide substitution. This analysis incorporates the genetic code and requires knowing which amino acid substitutions in the gene confer a scorable phenotype. The latter information is assembled by treating cells with a wide variety of mutagens, or by engineering a series of of mutations in the same gene. I am aware of only one such analysis, buried in (105). It provides a path for the rest of us to follow.

Materials and Methods

Datasets

UV datasets were identified by PubMed search using the search terms “ultraviolet”, “mutation”, and “sequence”; from papers listed in the references of these papers; and from papers listed in review articles. The list is intended to be complete and we apologize for any omissions. Datasets for H_2O_2, benzo(a)pyrene, and aflatoxin B_1 controls were identified similarly but without an attempt at completeness. Papers in which UV irradiation was carried out in medium rather than PBS, or which did not specify the irradiation conditions, were included but were flagged (83, 84). For counting mutations in mouse skin, we avoided counting mutant progeny cells by counting only one mutation of each type per mouse; this procedure will underestimate actual mutation frequency, particularly at hotspots, but this underestimate is standard in liquid culture mutagenesis experiments, where only one mutation of each type is counted per test tube. Datasets for skin lesions were limited to studies that microdissected lesions, amplified DNA (to avoid missing stop codon mutations due to nonsense-mediated mRNA decay), avoided a physical prescreening step such as mismatch detection, and examined nearly all of the *p53* gene.

Heat map

For comparison of mutation fraction data across all datasets, data was visualized using HeatMapViewer, v.13. Settings used were: Relative color scaling, color gradient. This program and all other software used are available at http://www.broadinstitute.org/cancer/software/genepattern/.

Hierarchical clustering

Initial clustering used the program Hierarchical Clustering, v.6 (106). Settings used were: Log transformed, row-mean subtracted. Column distance measure = Pearson correlation, Spearman correlation, or Kendall's tau. Clustering method = pairwise single-, complete- or centroid-linkage. Kendall's tau and pairwise-complete linkage showed the closest relation to treatment categories. Hierarchical clusters and their heat maps were visualized using Hierarchical Clustering Viewer, v.10.

K-Means clustering

To cluster by a different method, which also allows assessment of robustness by repeat clusterings of the same data using different random-number seeds, we performed K-means clustering (107) using the program K Means Clustering, v. 2. Settings used were: cluster by columns, Euclidean distance metric, 2 to 12 clusters. Here and for the NMF method, below,

missing data for the criterion “mutations at CG that are PyCG→PyTG” (at genes with few CG sites) were imputed with the average value for the dataset at CG-containing genes, 0.75.

Non-negative matrix factoring consensus clustering

To assess the robustness of clusters by their stability after multiple iterations, we used NMF consensus clustering (108). NMF has the advantage of clearly identifying datasets that are included in different clusters in different iterations and it tends to identify parts of a dataset that related to each other contextually (109). Data was analyzed using the program NMFConsensus, v.5. Settings used were: k initial = 2, k final = 5, 20 clusterings per value of k, 2000 iterations maximum, error function = divergence or Euclidean, iterations halted at 40 “no change” checks, “no change” checked each 10 iterations. Divergence showed less dispersion. Convergence was reached in <100 iterations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

For their insights on mutagenesis over the years, I heartily thank Evelyn Witkin, Jim Trosko, Veronica Maher, Eric Eisenstadt, Pat Foster, John Cairns, Graham Walker, Rick Bockrath, Barry Glickman, Roel Schaaper, Jan Drake, Ken Kraemer, Michael Seidman, Jim Cleaver, Larry Loeb, and Bill Thilly. These analyses were supported by a Developmental Research Project award to D.E.B from the Yale SPORE on Melanoma, NIH grant 5P50CA121974 to Ruth Halaban. I'm grateful to Drs. Christos Hatzis for advice on bioinformatics and Sanjay Premi for assistance with the figures. The term “canonical mutation” was suggested by Louise Perkins.

References

1. Angier, N. Ultraviolet radiation tied to gene defect producing skin cancer. 1991. Available at: http://www.nytimes.com/1991/11/19/science/ultraviolet-radiation-tied-to-genedefect-producing-skin-cancer.html
2. Brash DE, Rudolph JA, Simon JA, Lin A, McKenna GJ, Baden HP, Halperin AJ, Pontén J. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proc Natl Acad Sci USA. 1991; 88:10124–10128. [PubMed: 1946433]
3. Howard BD, Tessman I. Identification of the altered bases in mutated single-stranded DNA. 3. Mutagenesis by ultraviolet light. J Mol Biol. 1964; 9:372–375. [PubMed: 14202273]
4. Pettijohn D, Hanawalt P. Evidence for repair-replication of ultraviolet damaged DNA in bacteria. J Mol Biol. 1964; 9:395–410. [PubMed: 14202275]
5. Coulondre C, Miller JH. Genetic Studies of the *lac* Repressor IV. Mutagenic Specificity in the *lacI* Gene of *Escherichia coli*. J Mol Biol. 1977; 117:577–606. [PubMed: 416218]
6. Miller JH. Mutagenic specificity of ultraviolet light. J Mol Biol. 1985; 182:45–68. [PubMed: 3923204]
7. Brash DE, Haseltine WA. UV-induced mutation hotspots occur at DNA damage hotspots. Nature. 1982; 298:189–192. [PubMed: 7045692]
8. Glickman BW, Schaaper RM, Haseltine WA, Dunn RL, Brash DE. The C-C (6-4) UV photoproduct is mutagenic in *Escherichia coli*. Proc Natl Acad Sci U S A. 1986; 83:6945–6949. [PubMed: 3529093]
9. Brash DE, Seetharam S, Kraemer KH, Seidman MM, Bredberg A. Photoproduct frequency is not the major determinant of UV base substitution hot spots or cold spots in human cells. Proc Natl Acad Sci U S A. 1987; 84:3782–3786. [PubMed: 3473483]

10. Jans J, Schul W, Sert YG, Rijksen Y, Rebel H, Eker AP, Nakajima S, van Steeg H, de Gruijl FR, Yasui A, Hoeijmakers JH, van der Horst GT. Powerful skin cancer protection by a CPD-photolyase transgene. Curr Biol. 2005; 15:105–115. [PubMed: 15668165]
11. Pierceall WE, Mukhopadhyay T, Goldberg LH, Ananthaswamy HA. Mutations in the p53 tumor suppressor gene in human cutaneous squamous cell carcinoma. Mol Carcinog. 1991; 4:445–449. [PubMed: 1793482]
12. Rady P, Scinicariello F, Wagner RF, Tyring SK. p53 mutations in basal cell carcinomas. Cancer Res. 1992; 52:3804–3806. [PubMed: 1617650]
13. Campbell C, Quinn AG, Ro YS, Angus B, Rees JL. p53 mutations are common and early events that precede tumor invasion in squamous cell neoplasia of the skin. J Invest Dermatol. 1993; 100:746–748. [PubMed: 8496613]
14. Ziegler A, Leffell DJ, Kunala S, Sharma HW, Gailani M, Simon JA, Halperin AJ, Baden HP, Shapiro PE, Bale AE, Brash DE. Mutation hotspots due to sunlight in the p53 gene of non-melanoma skin cancers. Proc Natl Acad Sci U S A. 1993; 90:4216–4220. [PubMed: 8483937]
15. Ziegler A, Jonason AS, Leffell DJ, Simon JA, Sharma HW, Kimmelman J, Remington L, Jacks T, Brash DE. Sunburn and p53 in the onset of skin cancer. Nature. 1994; 372:773–776. [PubMed: 7997263]
16. Nelson MA, Einspahr JG, Alberts DS, Balfour CA, Wymer JA, Welch KL, Salasche SJ, Bangert JL, Grogan TM, Bozzo PO. Analysis of the *p53* gene in human precanerous actinic keratosis lesions and squamous cell cancers. Cancer Lett. 1994; 85:23–29. [PubMed: 7923098]
17. Taguchi M, Watanabe S, Yashima K, Murakami Y, Sekiya T, Ikeda S. Aberrations of the tumor suppressor p53 gene and p53 protein in solar keratosis in human skin. J Invest Dermatol. 1994; 103:500–503. [PubMed: 7930674]
18. Park WS, Lee HK, Lee JY, Yoo NJ, Kim CS, Kim SH. p53 mutations in solar keratoses. Hum Pathol. 1996; 27:1180–1184. [PubMed: 8912828]
19. Jonason AS, Kunala S, Price GJ, Restifo RJ, Spinelli HM, Persing JA, Leffell DJ, Tarone RE, Brash DE. Frequent clones of p53-mutated keratinocytes in normal human skin. Proc Natl Acad Sci USA. 1996; 93:14025–14029. [PubMed: 8943054]
20. Ren ZP, Ponten F, Nister M, Ponten J. Two distinct p53 immunohistochemical patterns in human squamous cell skin cancer, precursors, and normal epidermis. Int J Cancer. 1996; 69:174–179. [PubMed: 8682583]
21. Matsumura Y, Nishigori C, Yagi T, Imamura S, Takebe H. Characterization of p53 gene mutations in basal-cell carcinomas: comparison between sun-exposed and less-exposed skin areas. Int J Cancer. 1996; 65:778–780. [PubMed: 8631591]
22. Ratner D, Peacocke M, Zhang H, Ping XL, Tsou HC. UV-specific p53 and PTCH mutations in sporadic basal cell carcinoma of sun-exposed skin. J Am Acad Dermatol. 2001; 44:293–297. [PubMed: 11174390]
23. Kim MY, Park HJ, Baek SC, Byun DG, Houh D. Mutations of the p53 and PTCH gene in basal cell carcinomas: UV mutation signature and strand bias. J Dermatol Sci. 2002; 29:1–9. [PubMed: 12007715]
24. Weihrauch M, Bader M, Lehnert G, Wittekind C, Tannapfel A, Wrbitzky R. Carcinogen-specific mutation pattern in the p53 tumour suppressor gene in UV radiation-induced basal cell carcinoma. Int Arch Occup Environ Health. 2002; 75:272–276. [PubMed: 11981662]
25. Gailani MR, Stahle-Backdahl M, Leffell DJ, Glynn M, Zaphiropoulos PG, Pressman C, Unden AB, Dean M, Brash DE, Bale AE, Toftgard R. The role of the human homologue of *Drosophila patched* in sporadic basal cell carcinomas. Nat Genet. 1996; 14:78–81. [PubMed: 8782823]
26. Dumaz N, Drougard C, Sarasin A, Daya-Grosjean L. Specific UV-induced mutation spectrum in the p53 gene of skin tumors from DNA repair deficient xeroderma pigmentosum patients. Proc Natl Acad Sci USA. 1993; 90:10529–10533. [PubMed: 8248141]
27. Daya-Grosjean L, Sarasin A. The role of UV induced lesions in skin carcinogenesis: an overview of oncogene and tumor suppressor gene modifications in xeroderma pigmentosum skin tumors. Mutat Res. 2005; 571:43–56. [PubMed: 15748637]

28. Masaki T, Wang Y, DiGiovanna JJ, Khan SG, Raffeld M, Beltaifa S, Hornyak TJ, Darling TN, Lee CC, Kraemer KH. High frequency of PTEN mutations in nevi and melanomas from xeroderma pigmentosum patients. Pigment Cell Melanoma Res. 2014; 27:454–464. [PubMed: 24483290]

29. Pollock PM, Yu F, Parsons PG, Hayward NK. Evidence for u.v. induction of *CDKN2* mutations in melanoma cell lines. Oncogene. 1995; 11:663–668. [PubMed: 7651729]

30. Krauthammer M, Kong Y, Ha BH, Evans P, Bacchiocchi A, McCusker JP, Cheng E, Davis MJ, Goh G, Choi M, Ariyan S, Narayan D, Dutton-Regester K, Capatana A, Holman EC, Bosenberg M, Sznol M, Kluger HM, Brash DE, Stern DF, Materin MA, Lo RS, Mane S, Ma S, Kidd KK, Hayward NK, Lifton RP, Schlessinger J, Boggon TJ, Halaban R. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat Genet. 2012; 44:1006–1014. [PubMed: 22842228]

31. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, Dicara D, Ramos AH, Lawrence MS, Cibulskis K, Sivachenko A, Voet D, Saksena G, Stransky N, Onofrio RC, Winckler W, Ardlie K, Wagle N, Wargo J, Chong K, Morton DL, Stemke-Hale K, Chen G, Noble M, Meyerson M, Ladbury JE, Davies MA, Gershenwald JE, Wagner SN, Hoon DS, Schadendorf D, Lander ES, Gabriel SB, Getz G, Garraway LA, Chin L. A landscape of driver mutations in melanoma. Cell. 2012; 150:251–263. [PubMed: 22817889]

32. Greenblatt MS, Bennett WP, Hollstein M, Harris CC. Mutations in the *p53* tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Res. 1994; 54:4855–4878. [PubMed: 8069852]

33. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, I Australian Pancreatic Cancer Genome; I B C. Consortium; I M S Consortium. PedBrain I, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, Stratton MR. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–421. [PubMed: 23945592]

34. Drobetsky EA, Moustacchi E, Glickman BW, Sage E. The mutational specificity of simulated sunlight at the aprt locus in rodent cells. Carcinogenesis. 1994; 15:1577–1583. [PubMed: 8055636]

35. Morreall JF, Petrova L, Doetsch PW. Transcriptional mutagenesis and its potential roles in the etiology of cancer and bacterial antibiotic resistance. J Cell Physiol. 2013; 228:2257–2261. [PubMed: 23696333]

36. Kim JY, Tavare S, Shibata D. Counting human somatic cell replications: methylation mirrors endometrial stem cell divisions. Proc Natl Acad Sci U S A. 2005; 102:17739–17744. [PubMed: 16314580]

37. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A. 2011; 108:9530–9535. [PubMed: 21586637]

38. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A. 2012; 109:14508–14513. [PubMed: 22853953]

39. Reid TM, Loeb LA. Tandem double CC-->TT mutations are produced by reactive oxygen species. Proc Natl Acad Sci U S A. 1993; 90:3904–3907. [PubMed: 8483909]

40. Newcomb TG, Allen KJ, Tkeshelashvili L, Loeb LA. Detection of tandem CC-->TT mutations induced by oxygen radicals using mutation-specific PCR. Mutat Res. 1999; 427:21–30. [PubMed: 10354498]

41. Agar NS, Halliday GM, Barnetson RS, Ananthaswamy HN, Wheeler M, Jones AM. The basal layer in human squamous tumors harbors more UVA than UVB fingerprint mutations: a role for

UVA in human skin carcinogenesis. Proc Natl Acad Sci U S A. 2004; 101:4954–4959. [PubMed: 15041750]

42. Thomas NE, Berwick M, Cordeiro-Stone M. Could BRAF mutations in melanocytic lesions arise from DNA damage induced by ultraviolet radiation? J Invest Dermatol. 2006; 126:1693–1696. [PubMed: 16845408]
43. Sage E, Drobetsky EA, Moustacchi E. 8-Methoxypsoralen induced mutations are highly targeted at crosslinkable sites of photoaddition on the non-transcribed strand of a mammalian chromosomal gene. EMBO J. 1993; 12:397–402. [PubMed: 8440233]
44. Hussain SP, Schwank J, Staib F, Wang XW, Harris CC. TP53 mutations and hepatocellular carcinoma: insights into the etiology and pathogenesis of liver cancer. Oncogene. 2007; 26:2166–2176. [PubMed: 17401425]
45. Ikehata H, Ono T. Significance of CpG methylation for solar UV-induced mutagenesis and carcinogenesis in skin. Photochem Photobiol. 2007; 83:196–204. [PubMed: 16620158]
46. Runger TM, Kappes UP. Mechanisms of mutation formation with long-wave ultraviolet light (UVA). Photodermatol Photoimmunol Photomed. 2008; 24:2–10. [PubMed: 18201350]
47. Ikehata H, Ono T. The mechanisms of UV mutagenesis. J Radiat Res. 2011; 52:115–125. [PubMed: 21436607]
48. Pfeifer GP, Besaratinia A. UV wavelength-dependent DNA damage and human non-melanoma and melanoma skin cancer. Photochem Photobiol Sci. 2012; 11:90–97. [PubMed: 21804977]
49. Sage E, Girard PM, Francesconi S. Unravelling UVA-induced mutagenesis. Photochem Photobiol Sci. 2012; 11:74–80. [PubMed: 21901217]
50. Lebkowski JS, Clancy S, Miller JH, Calos MP. The *lacI* shuttle: rapid analysis of the mutagenic specificity of ultraviolet light in human cells. Proc Natl Acad Sci USA. 1985; 82:8606–8610. [PubMed: 3001711]
51. Urlaub G, Mitchell PJ, Ciudad CJ, Chasin LA. Nonsense mutations in the dihydrofolate reductase gene affect RNA processing. Mol Cell Biol. 1989; 9:2868–2880. [PubMed: 2779551]
52. Glazer PM, Sarkar SN, Summers WC. Detection and analysis of UV-induced mutations in mammalian cell DNA using a lambda phage shuttle vector. Proc Natl Acad Sci USA. 1986; 83:1041–1044. [PubMed: 2937054]
53. Oller AR, Thilly WG. Mutational spectra in human B-cells. Spontaneous, oxygen and hydrogen peroxide-induced mutations at the hprt gene. J Mol Biol. 1992; 228:813–826. [PubMed: 1469715]
54. Persson AE, Edstrom DW, Backvall H, Lundeberg J, Ponten F, Ros AM, Williams C. The mutagenic effect of ultraviolet-A1 on human skin demonstrated by sequencing the p53 gene in single keratinocytes. Photodermatol Photoimmunol Photomed. 2002; 18:287–293. [PubMed: 12535024]
55. Huang XX, Bernerd F, Halliday GM. Ultraviolet A within sunlight induces mutations in the epidermal basal layer of engineered human skin. Am J Pathol. 2009; 174:1534–1543. [PubMed: 19264911]
56. Hauser J, Seidman MM, Sidur K, Dixon K. Sequence specificity of point mutations induced during passage of a UV-irradiated shuttle vector plasmid in monkey cells. Mol Cell Biol. 1986; 6:277–285. [PubMed: 3537686]
57. Bredberg A, Kraemer KH, Seidman MM. Restricted ultraviolet mutational spectrum in a shuttle vector propagated in xeroderma pigmentosum cells. Proc Natl Acad Sci USA. 1986; 83:8273–8277. [PubMed: 3464953]
58. Seetharam S, Protic-Sabljic M, Seidman MM, Kraemer KH. Abnormal ultraviolet mutagenic spectrum in plasmid DNA replicated in cultured fibroblasts from a patient with the skin cancer-prone disease, xeroderma pigmentosum. J Clin Invest. 1987; 80:1613–1617. [PubMed: 3680516]
59. Drobetsky EA, Grosovsky AJ, Glickman BW. The specificity of UV-induced mutations at an endogenous locus in mammalian cells. Proc Natl Acad Sci USA. 1987; 84:9103–9107. [PubMed: 3480533]
60. Keyse SM, Amaudruz F, Tyrrell RM. Determination of the spectrum of mutations induced by defined-wavelength solar UVB (313 nm) radiation in mammalian cells by use of a shuttle vector. Mol Cell Biol. 1988; 8:5425–5431. [PubMed: 3072480]

61. Hsia HC, Lebkowski JS, Leong PM, Calos MP, Miller JH. Comparison of ultraviolet irradiation-induced mutagenesis of the lacI gene in Escherichia coli and in human 293 cells. J Mol Biol. 1989; 205:103–113. [PubMed: 2647996]

62. Drobetsky EA, Grosovsky AJ, Skandalis A, Glickman BW. Perspectives on UV light mutagenesis: investigation of the CHO aprt gene carried on a retroviral shuttle vector. Somat Cell Mol Genet. 1989; 15:401–409. [PubMed: 2781414]

63. Romac S, Leong P, Sockett H, Hutchinson F. DNA base sequence changes induced by ultraviolet light mutagenesis of a gene on a chromosome in Chinese hamster ovary cells. J Mol Biol. 1989; 209:195–204. [PubMed: 2685319]

64. Vrieling H, van Rooijen ML, Groen NA, Zdzienicka MZ, Simons JWIM, Lohman PHM, van Zeeland AA. DNA strand specificity for UV-induced mutations in mammalian cells. Mol Cell Biol. 1989; 9:1277–1283. [PubMed: 2725498]

65. Drobetsky EA, Glickman BW. The nature of ultraviolet light-induced mutations at the heterozygous aprt locus in Chinese hamster ovary cells. Mutat Res. 1990; 232:281–289. [PubMed: 1977078]

66. Dorado G, Steingrimsdottir H, Arlett CF, Lehmann AR. Molecular analysis of ultraviolet-induced mutations in a xeroderma pigmentosum cell line. J Mol Biol. 1991; 217:217–222. [PubMed: 1992158]

67. McGregor WG, Chen RH, Lukash L, Maher VM, McCormick JJ. Cell cycle-dependent strand bias for UV-induced mutations in the transcribed strand of excision repair-proficient human fibroblasts but not in repair-deficient cells. Mol Cell Biol. 1991; 11:1927–1934. [PubMed: 2005888]

68. Sage E, Lamolet B, Brulay E, Moustacchi E, Chateauneuf A, Drobetsky EA. Mutagenic specificity of solar UV light in nucleotide excision repair-deficient rodent cells. Proc Natl Acad Sci USA. 1996; 93:176–180. [PubMed: 8552599]

69. You YH, Li C, Pfeifer GP. Involvement of 5-methylcytosine in sunlight-induced mutagenesis. J Mol Biol. 1999; 293:493–503. [PubMed: 10543945]

70. You YH, Pfeifer GP. Similarities in sunlight-induced mutational spectra of CpG-methylated transgenes and the p53 gene in skin cancer point to an important role of 5-methylcytosine residues in solar UV mutagenesis. J Mol Biol. 2001; 305:389–399. [PubMed: 11152598]

71. Borgdorff V, Pauw B, van Hees-Stuivenberg S, de Wind N. DNA mismatch repair mediates protection from mutagenesis induced by short-wave ultraviolet light. DNA Repair (Amst). 2006; 5:1364–1372. [PubMed: 16880010]

72. Vreeswijk MP, Meijers CM, Giphart-Gassler M, Vrieling H, van Zeeland AA, Mullenders LH, Loenen WA. Site-specific analysis of UV-induced cyclobutane pyrimidine dimers in nucleotide excision repair-proficient and -deficient hamster cells: Lack of correlation with mutational spectra. Mutat Res. 2009; 663:7–14. [PubMed: 19150617]

73. Drobetsky EA, Turcotte J, Chateauneuf A. A role for UVA in solar mutagenesis. Proc Natl Acad Sci USA. 1995; 92:2350–2354. [PubMed: 7892270]

74. Robert C, Muel B, Benoit A, Dubertret L, Sarasin A, Stary A. Cell survival and shuttle vector mutagenesis induced by ultraviolet A and ultraviolet B radiation in a human cell line. J Invest Dermatol. 1996; 106:721–728. [PubMed: 8618011]

75. Frijhoff AF, Rebel H, Mientjes EJ, Kelders MC, Steenwinkel MJ, Baan RA, van Zeeland AA, Roza L. UVB-induced mutagenesis in hairless lambda lacZ-transgenic mice. Environ Mol Mutagen. 1997; 29:136–142. [PubMed: 9118965]

76. Horiguchi M, Masumura K, Ikehata H, Ono T, Kanke Y, Sofuni T, Nohmi T. UVB-induced gpt mutations in the skin of gpt delta transgenic mice. Environ Mol Mutagen. 1999; 34:72–79. [PubMed: 10529728]

77. Murai H, Takeuchi S, Nakatsu Y, Ichikawa M, Yoshino M, Gondo Y, Katsuki M, Tanaka K. Studies of in vivo mutations in rpsL transgene in UVB-irradiated epidermis of XPA-deficient mice. Mutat Res. 2000; 450:181–192. [PubMed: 10838142]

78. You YH, Lee DH, Yoon JH, Nakajima S, Yasui A, Pfeifer GP. Cyclobutane pyrimidine dimers are responsible for the vast majority of mutations induced by UVB irradiation in mammalian cells. J Biol Chem. 2001; 276:44688–44694. [PubMed: 11572873]

79. Ikehata H, Masuda T, Sakata H, Ono T. Analysis of mutation spectra in UV-Bexposed mouse skin epidermis and dermis: frequent occurrence of C-->T transition at methylated CpG-associated dipyrimidine sites. Environ Mol Mutagen. 2003; 41:280–292. [PubMed: 12717783]

80. Kappes UP, Luo D, Potter M, Schulmeister K, Runger TM. Short- and long-wave UV light (UVB and UVA) induce similar mutations in human skin cells. J Invest Dermatol. 2006; 126:667–675. [PubMed: 16374481]

81. Ikehata H, Kudo H, Masuda T, Ono T. UVA induces C-->T transitions at methyl-CpG-associated dipyrimidine sites in mouse skin epidermis more frequently than UVB. Mutagenesis. 2003; 18:511–519. [PubMed: 14614186]

82. Besaratinia A, Bates SE, Synold TW, Pfeifer GP. Similar mutagenicity of photoactivated porphyrins and ultraviolet A radiation in mouse embryonic fibroblasts: involvement of oxidative DNA lesions in mutagenesis. Biochemistry. 2004; 43:15557–15566. [PubMed: 15581368]

83. Besaratinia A, Synold TW, Xi B, Pfeifer GP. G-to-T transversions and small tandem base deletions are the hallmark of mutations induced by ultraviolet a radiation in mammalian cells. Biochemistry. 2004; 43:8169–8177. [PubMed: 15209513]

84. Kim SI, Pfeifer GP, Besaratinia A. Mutagenicity of ultraviolet A radiation in the lacI transgene in Big Blue mouse embryonic fibroblasts. Mutat Res. 2007; 617:71–78. [PubMed: 17275039]

85. Besaratinia A, Kim SI, Bates SE, Pfeifer GP. Riboflavin activated by ultraviolet A1 irradiation induces oxidative DNA damage-mediated mutations inhibited by vitamin C. Proc Natl Acad Sci U S A. 2007; 104:5953–5958. [PubMed: 17389394]

86. Ikehata H, Kawai K, Komura J, Sakatsume K, Wang L, Imai M, Higashi S, Nikaido O, Yamamoto K, Hieda K, Watanabe M, Kasai H, Ono T. UVA1 genotoxicity is mediated not by oxidative damage but by cyclobutane pyrimidine dimers in normal mouse skin. J Invest Dermatol. 2008; 128:2289–2296. [PubMed: 18356809]

87. Ikehata H, Kumagai J, Ono T, Morita A. Solar-UV-signature mutation prefers TCG to CCG: extrapolative consideration from UVA1-induced mutation spectra in mouse skin. Photochem Photobiol Sci. 2013; 12:1319–1327. [PubMed: 23471200]

88. Ikehata H, Nakamura S, Asamura T, Ono T. Mutation spectrum in sunlight-exposed mouse skin epidermis: small but appreciable contribution of oxidative stressmediated mutagenesis. Mutat Res. 2004; 556:11–24. [PubMed: 15491628]

89. Yang JL, Chen RH, Maher VM, McCormick JJ. Kinds and location of mutations induced by (+/-)-7 beta,8 alpha-dihydroxy-9 alpha,10 alpha-epoxy-7,8,9,10-tetrahydrobenzo[a]pyrene in the coding region of the hypoxanthine (guanine) phosphoribosyltransferase gene in diploid human fibroblasts. Carcinogenesis. 1991; 12:71–75. [PubMed: 1899056]

90. Levy DD, Groopman JD, Lim SE, Seidman MM, Kraemer KH. Sequence specificity of aflatoxin B_1-induced mutations in a plasmid replicated in xeroderma pigmentosum and DNA repair proficient human cells. Cancer Res. 1992; 52:5668–5673. [PubMed: 1394191]

91. Trottier Y, Waithe WI, Anderson A. Kinds of mutations induced by aflatoxin B1 in a shuttle vector replicating in human cells transiently expressing cytochrome P4501A2 cDNA. Mol Carcinog. 1992; 6:140–147. [PubMed: 1326989]

92. Courtemanche C, Anderson A. Shuttle-vector mutagenesis by aflatoxin B1 in human cells: effects of sequence context on the supF mutational spectrum. Mutat Res. 1994; 306:143–151. [PubMed: 7512213]

93. Dycaico MJ, Stuart GR, Tobal GM, de Boer JG, Glickman BW, Provost GS. Species-specific differences in hepatic mutant frequency and mutational spectrum among lambda/lacI transgenic rats and mice following exposure to aflatoxin B1. Carcinogenesis. 1996; 17:2347–2356. [PubMed: 8968048]

94. Diaz-Llera S, Podlutsky A, Osterholm AM, Hou SM, Lambert B. Hydrogen peroxide induced mutations at the HPRT locus in primary human Tlymphocytes. Mutat Res. 2000; 469:51–61. [PubMed: 10946242]

95. Bastien N, Therrien JP, Drouin R. Cytosine containing dipyrimidine sites can be hotspots of cyclobutane pyrimidine dimer formation after UVB exposure. Photochem Photobiol Sci. 2013; 12:1544–1554. [PubMed: 23877442]

96. Drouin R, Therrien JP. UVB-induced cyclobutane pyrimidine dimer frequency correlates with skin cancer mutational hotspots in p53. Photochem Photobiol. 1997; 66:719–726. [PubMed: 9383997]

97. Tommasi S, Denissenko MF, Pfeifer GP. Sunlight induces pyrimidine dimers preferentially at 5-methylcytosine bases. Cancer Res. 1997; 57:4727–4730. [PubMed: 9354431]

98. Setlow RB, Carrier WL. Pyrimidine dimers in ultraviolet-irradiated DNA's. J Mol Biol. 1966; 17:237–254. [PubMed: 4289765]

99. Fix D, Bockrath R. Thermal resistance to photoreactivation of specific mutations potentiated in E. coli B/r ung by ultraviolet light. Mol Gen Genet : MGG. 1981; 182:7–11.

100. Jiang N, Taylor JS. In vivo evidence that UV-induced C-->T mutations at dipyrimidine sites could result from the replicative bypass of cis-syn cyclobutane dimers or their deamination products. Biochemistry. 1993; 32:472–481. [PubMed: 8422356]

101. Peng W, Shaw BR. Accelerated deamination of cytosine residues in UV-induced cyclobutane pyrimidine dimers leads to CC-->TT transitions. Biochemistry. 1996; 35:10172–10181. [PubMed: 8756482]

102. Tu Y, Dammann R, Pfeifer GP. Sequence and time-dependent deamination of cytosine bases in UVB-induced cyclobutane pyrimidine dimers in vivo. J Mol Biol. 1998; 284:297–311. [PubMed: 9813119]

103. Cannistraro VJ, Taylor JS. Acceleration of 5-methylcytosine deamination in cyclobutane dimers by G and its implications for UV-induced C-to-T mutation hotspots. J Mol Biol. 2009; 392:1145–1157. [PubMed: 19631218]

104. Baserga SJ, Benz EJ. Nonsense mutations in the human beta-globin gene affect mRNA metabolism. Proc Natl Acad Sci USA. 1988; 85:2056–2060. [PubMed: 3353367]

105. Rodin SN, Rodin AS, Juhasz A, Holmquist GP. Cancerous hyper-mutagenesis in p53 genes is possibly associated with transcriptional bypass of DNA lesions. Mutat Res. 2002; 510:153–168. [PubMed: 12459451]

106. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95:14863–14868. [PubMed: 9843981]

107. MacQueen, JB. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1967; University of California Press; p. 281-297.

108. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A. 2004; 101:4164–4169. [PubMed: 15016911]

109. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401:788–791. [PubMed: 10548103]

Biography

Douglas E. Brash is Professor of Therapeutic Radiology and Dermatology at Yale University. After receiving degrees in Engineering Physics and Biophysics, he joined the William Haseltine laboratory at Harvard to apply DNA sequencing techniques to DNA repair and mutagenesis. At the National Cancer Institute and then Yale, his laboratory identified DNA photoproducts responsible for UV mutations, used UV signatures to identify genes mutated by sunlight in the course of generating skin cancers, showed that one of these genes, *p53*, is needed for UV-induced apoptosis to remove cells that otherwise would lead to cancer, and that this apoptosis also drives clonal expansion of mutant cells once they arise. Current interests include pro-cancerous effects of melanin and the role of UV in tumor evolution.

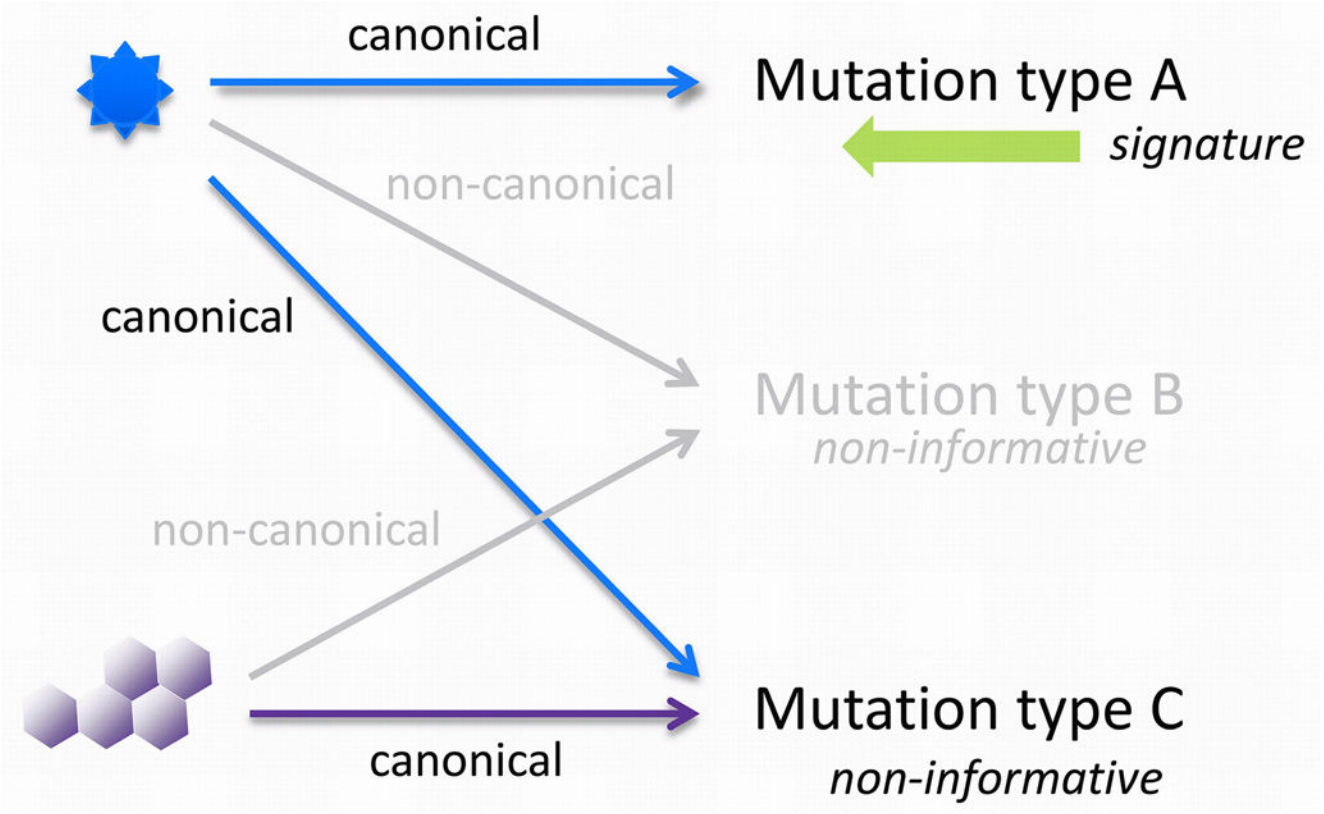

Figure 1.

Inverse relationship of canonical mutation patterns and mutation signatures for inferring the mutagen from mutations. Two mutagens are illustrated. A mutagen's canonical mutations deviate from random base changes, establishing a pattern typical for that mutagen. Different mutagens can produce the same canonical mutations (non-informative mutations). Signature mutations are the subset of canonical mutations that, in addition, are unique to that mutagen and permit inference backward from mutations to mutagen. A mutagen therefore produces signature mutations plus non-informative mutations. The latter are real and were produced by the mutagen, but are not useful for identifying that mutagen or carcinogen.

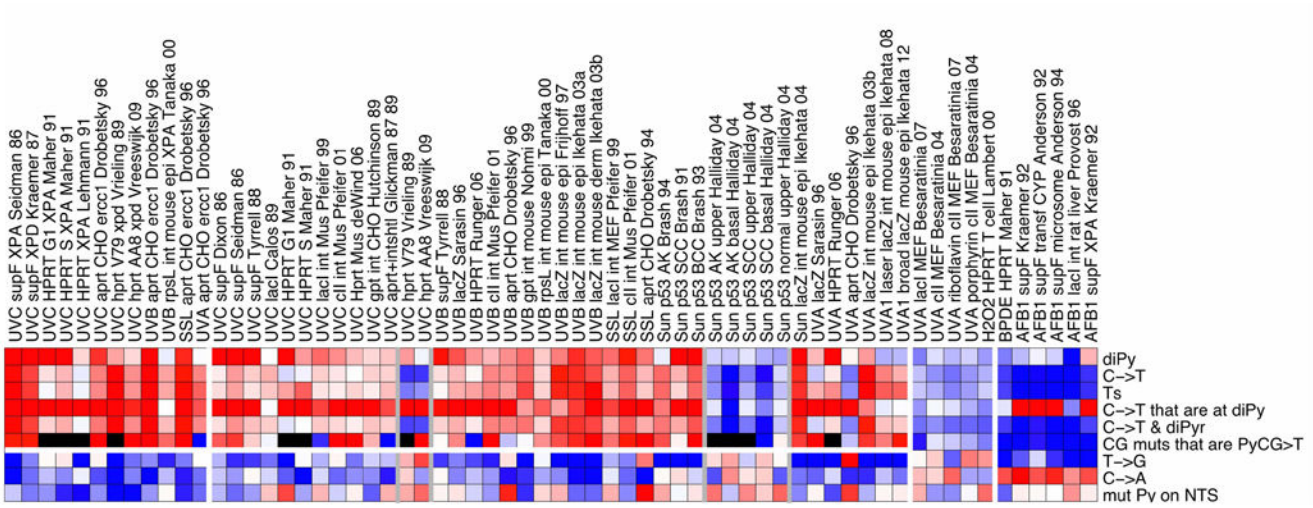

Figure 2.

Heat map of canonical mutation patterns after UVC, UVB, UVA, SSL, and chemical carcinogens. When many gene targets are analyzed, the canonical mutation patterns of all four UV wavelength regions are similar. Yet they differ from the chemical mutagens. Finer patterns are discussed in the text. Colors: dark blue, row minimum; white, row average; dark red, row maximum. Lines, distinctive subsets. Abbreviations: diPY, the mutated base was a member of a dipyrimidine site; C->T, C → T mutation (G:C → A:T); Ts, transition mutation (Py:Pu → Py:Pu); CG muts, the mutated base was a member of a CG dinucleotide; NTS, nontranscribed strand.

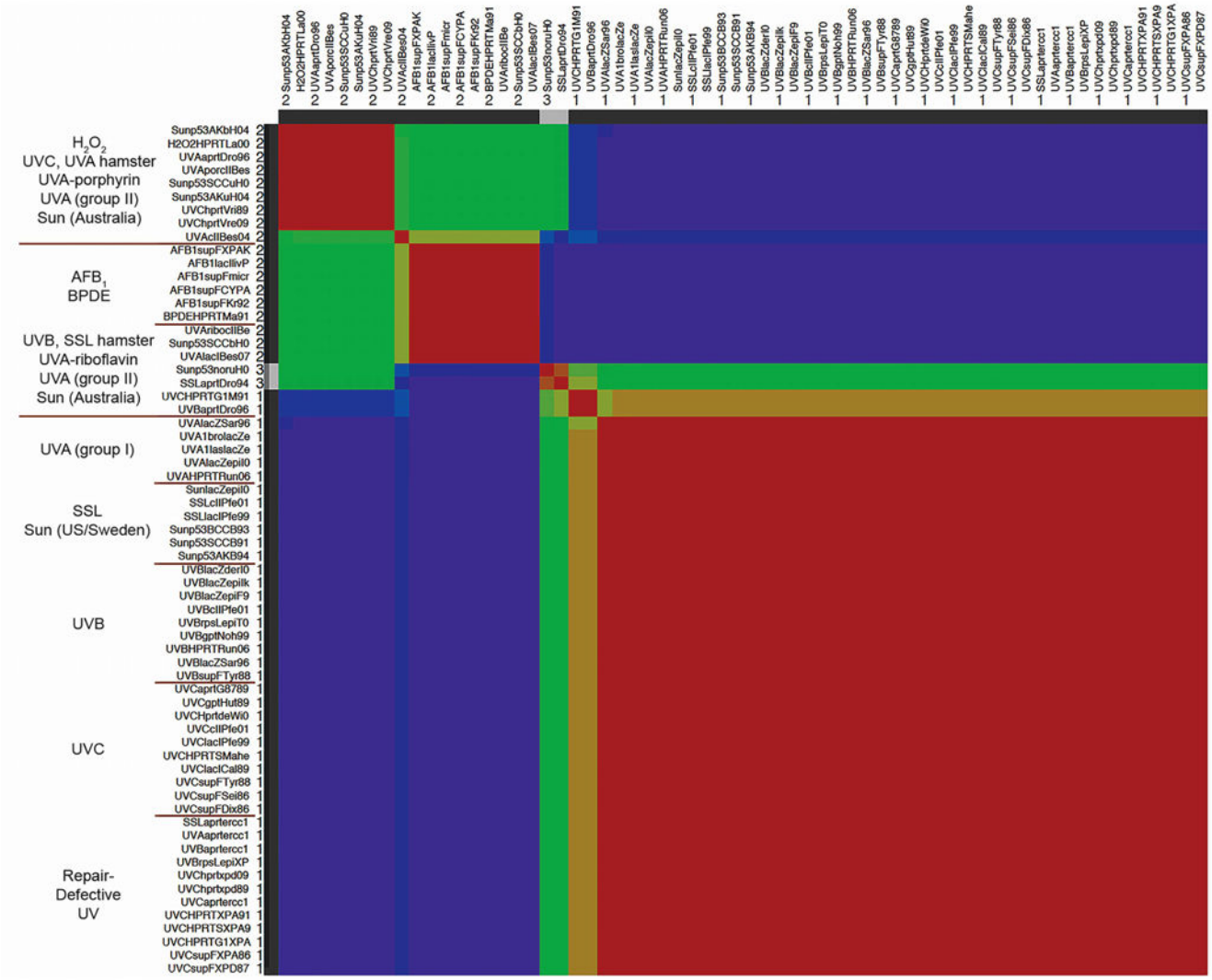

Figure 3.

Clustering of canonical mutation patterns by non-negative matrix factorization (NMF). Colors indicate the degree of correlation between datasets listed on both the vertical and horizontal axes. Two to three stable groups were identifiable (dark red): a) repair-defective cells (any UV wavelength) + wild type UVC + UVB + SSL + Sun (US/Sweden) + UVA group I; b) Sun (Australia) + UVA group II + UVA-riboflavin + UVB&SSL hamster + BPDE + AFB_1; and c) Sun (Australia) + UVA group II + UVA-porphyrin + UVC&UVA hamster + H_2O_2. Clusters were separated from each other by vertical and horizontal bands representing datasets that clustered into one or the other of its neighbors on different runs (green, orange, and light blue bands). Abbreviations indicate a particular report's mutagen, gene target, author, and year; these are written in full in Table S1.

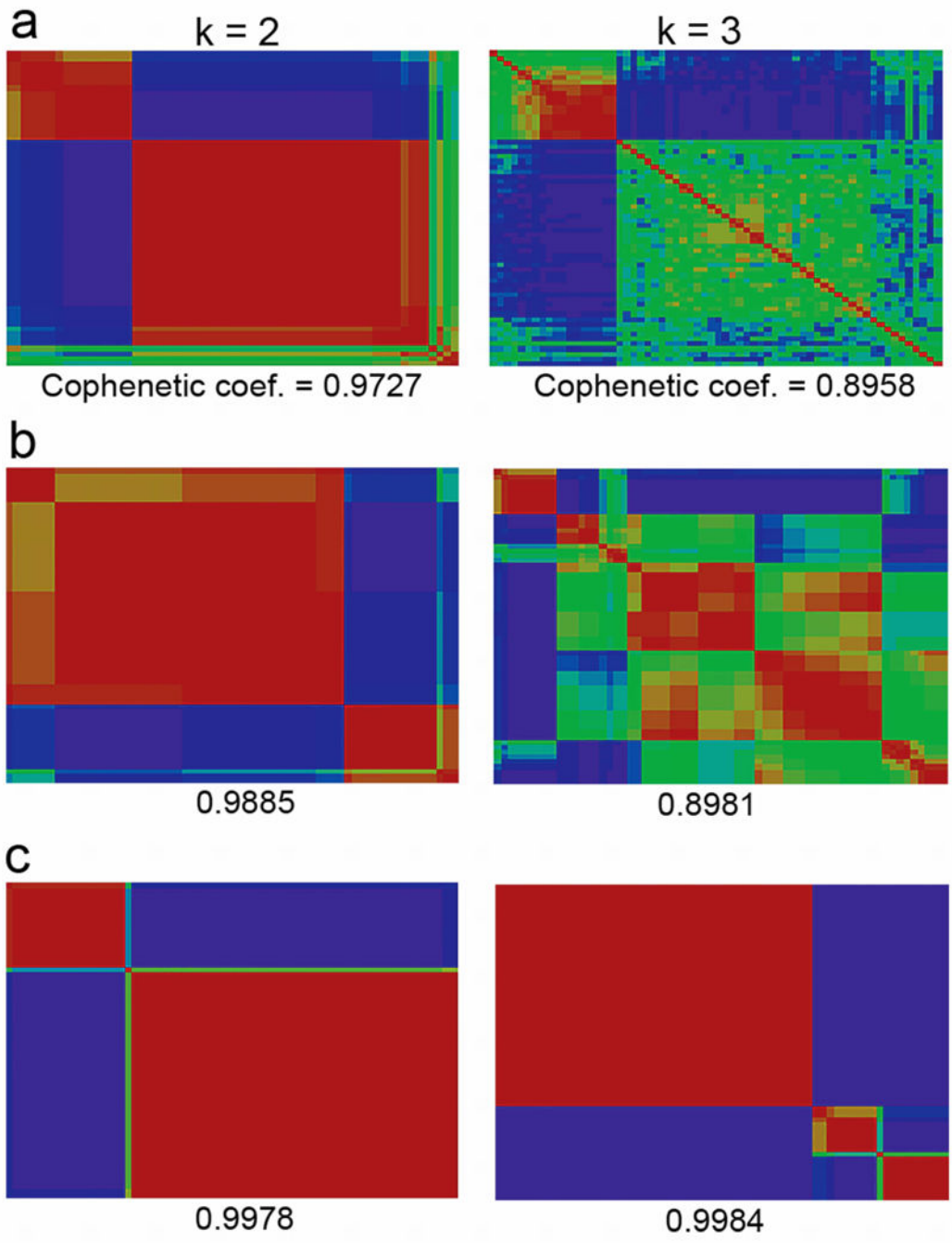

Figure 4.

Discrimination of UV-induced mutations from chemically-induced mutations is enhanced by including negative-control mutation features. a) Using only the 2 canonical UV features, "dipyrimidine site" and "C→T", yielded clusters that showed much greater dispersion than in Fig. 3 and it intermixed datasets that had been separated in the NMF analysis when using 9 features. Dispersion is apparent visually and as a lower cophenetic score. Dispersion was even greater when three clusters were stipulated at the outset. b) Adding 2 additional UVlike features, "transition mutations" and "CG mutations that are at PyCG>T", improved

dispersion slightly. c) Much greater resolution, revealing three clusters, was obtained by instead supplementing the 2 original canonical UV features with 2 negative controls: features rare with UV (A:T→C:G and G:C→T:A) (lower row). These are non-signature canonical UV mutations and non-UV canonical mutations.