Hum Genet (2015) 134:865–867
DOI 10.1007/s00439-015-1576-z

ERRATUM

Erratum to: The somatic autosomal mutation matrix in cancer genomes

Nuri A. Temiz[1,4] · Duncan E. Donohue[1,5] · Albino Bacolla[1,2] · Karen M. Vasquez[2] · David N. Cooper[3] · Uma Mudunuri[1] · Joseph Ivanic[1] · Regina Z. Cer[1,6] · Ming Yi[1] · Robert M. Stephens[1] · Jack R. Collins[1] · Brian T. Luke[1]

Published online: 13 June 2015

Erratum to: Hum Genet DOI 10.1007/s00439-015-1566-1

In reviewing the software used in this publication, we detected an error in the distance-dependent K-nearest neighbor algorithm. Correction of the algorithm increased the accuracy of the tissue of origin with the distance-dependent 6-nearest neighbor classifier. The second to last sentence in the abstract should read:

"When a distance-dependent 6-nearest neighbor classifier was used, 10.4 % of the SAMMs had an Undetermined tissue of origin, and 92.2 % of the remaining SAMMs were assigned to the correct tissue of origin".

This correction also affects Supplemental Tables 5b through 5k, and Supplemental Table 6. The revised Discussion on inferring the tissue of origin and conclusion are as follows.

Inferring the cancer tissue of origin

Although some tissue specificity was evident for each of the four mutational mechanisms, the majority of all mutation frequencies were not accounted for by any of these mutational mechanisms. Thus, it is possible that at least some of these "orphan" mutation frequency patterns are tissue or cell-type specific. Mutation patterns, along with clinical, transcriptional and other data, have been integrated to improve the classification of tumor subtypes (Hoadley et al. 2014; Kandoth et al. 2013). We have used our Manhattan distances to place all 908 SAMMs (excluding the single ALL cancer genome) in relation to one another to assess the extent to which the correct tissue of origin of a tumor can be inferred from its position relative to all tumor samples, based on the identity of its nearest neighbor. For example, a breast tumor SAMM would be identified as having the correct tissue of origin if the nearest neighbor were to belong to either of the breast cancer datasets. The same would be true for the three liver cancer datasets as well as the four pancreatic cancer datasets. Samples from the medulloblastoma and pediatric brain cancer datasets were also assigned to the same tissue of origin. The results

The online version of the original article can be found under doi:10.1007/s00439-015-1566-1.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-015-1576-z) contains supplementary material, which is available to authorized users.

✉ Brian T. Luke
Brian.Luke@nih.gov

1 In Silico Research Centers of Excellence, Advanced Biomedical Computing Center, Information Systems Program, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research Inc., P.O. Box B, Frederick, MD 21702, USA

2 Division of Pharmacology and Toxicology, The University of Texas at Austin, Austin, TX 78723, USA

3 Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK

4 Present Address: Masonic Cancer Center, University of Minnesota, 2-120 CCRB, 2231 6th St SE, Minneapolis, MN 55455, USA

5 Present Address: US Army Medical Research and Material Command, 568 Doughten Dr., Fort Detrick, Frederick, MD 21702, USA

6 Present Address: Naval Medical Research Center-Frederick, 8400 Research Plaza, Fort Detrick, Frederick, MD 21702, USA

are reported in Supplemental Table 4, which lists each sample and its dataset, as well as the nearest neighbor sample, its dataset, and the distance between their SAMMs. We were gratified that our procedure was able to identify the correct tissue of origin 89.5 % of the time, a finding which supports the emerging view that cell-type-specific mutational processes are prevalent in cancer biology (Hoadley et al. 2014).

Extending our analysis further, a distance-dependent *k*-nearest neighbor classifier was used to predict the tissue of origin. Along with the single ALL cancer genome, the four samples in the LAML-KR dataset were excluded from this analysis due to an insufficient number of samples. The remaining 904 cancer genomes were placed into 11 Groups, or cancer types, as shown in Supplemental Table 5a. The initial analysis used six nearest neighbors ($k = 6$) and the un-normalized probability of belonging to a neighbor's Group was 0.5 when the Manhattan distance to this neighbor was 0.5 ($D_{0.5} = 0.5$, $\beta = 0.25$). The value for $D_{0.5}$ appeared reasonable since, overall, the average distance to the nearest neighbor SAMM was 0.143, with a standard deviation of 0.065.

Once the probabilities were scaled to a sum of 1.0, each SAMM was assigned to a particular Group if the probability of belonging to that Group was at least 0.5; otherwise, the predicted tissue of origin was Undetermined. The results for each of the 11 Groups is shown in Supplemental Table 5b, and the distribution of probabilities across all 11 Groups and the Undetermined Group is shown for each sample in Supplemental Table 6. In total, 92.2 % of all samples were assigned to the correct tissue of origin from a comparison of those correctly classified and those assigned to the wrong tissue of origin, and 10.4 % of all samples had an Undetermined tissue of origin.

In renal cell carcinomas, 2 of the 71 SAMMs (2.8 %) received an Undetermined classification, and 68 of the remaining 69 (98.6 %) were correctly classified (Supplemental Table 5b). DO46877 was predicted to be a pancreatic cancer (Supplemental Table 6). 295 of the 303 liver tumors were assigned to the correct tissue of origin; 3 were assigned to the wrong tissue (pancreatic) and 5 were Undetermined. Conversely, only one of the ten prostate samples was assigned the correct tissue of origin; three were assigned to the wrong tissue and six were Undetermined. Overall, 6 of the 11 tissue types had a classification accuracy above 92 %, though 10.4 % of all samples received an Undetermined classification.

To reduce the number of SAMMs with an Undetermined classification, a maximum likelihood procedure was used, where each SAMM was assigned to the cancer tissue type with the highest probability, independent of its value. Following this analysis, no sample received an Undetermined classification (Supplemental Table 5c). Overall, 87.6 % of all samples (792 of the 904 SAMMs) were assigned to the correct tissue of origin. Only 4 of the 10 prostate SAMMs were assigned to the correct tissue of origin, while 298 of the 303 liver SAMMs were correctly classified.

To determine the effect of the classification accuracy on the number of nearest neighbors (k), the classifications were performed with k varying from three to eight. If the final assignment was made by requiring the scaled probability to be at least 0.5, between 3.2 and 10.4 % of all samples were Undetermined, and between 90.9 and 92.8 % of the remaining SAMMs yielded the correct tissue of origin (Supplemental Table 5d). When the maximum likelihood criterion was used, between 87.5 and 89.7 % of the 904 SAMMs were correctly classified (Supplemental Table 5e), with no Undetermined classifications.

To examine the effect of varying $D_{0.5}$ on the classification accuracy, a distance-dependent 6-nearest neighbor classifier was used with $D_{0.5}$ varying from 0.4 to 0.7. When the scaled probability is required to be at least 0.5 (Supplemental Table 5f), between 9.7 and 10.6 % of the SAMMs had an Undetermined tissue of origin and between 91.8 and 92.3 % of the remaining samples were assigned to the correct tissue of origin. The maximum likelihood procedure yielded an 87.6 % correct classification with no samples Undetermined for all values of $D_{0.5}$ (Supplemental Table 5g).

Overall, these results suggest that the classification accuracy for the tissue of origin is relatively insensitive to the number of nearest neighbors and the value of $D_{0.5}$. While the maximum likelihood criterion reduced the number of Undetermined tissues of origin, forcing these outliers into one of the 11 tissue Groups led to a larger number of incorrect classifications, thereby reducing the percentage of SAMMs assigned to the correct tissue of origin.

To compare these results with a standard k-nearest neighbor classifier that does not use $D_{0.5}$ and P(und), the classification of tissue of origin was repeated varying k from three to eight. Requiring a scaled probability of at least 0.5 for a classification (Supplemental Table 5h), between 0.6 and 6.2 % of the samples had an Undetermined tissue of origin. The highest level was for 7-nearest neighbors, where for 56 SAMMs their 7 neighboring SAMMs were heterogeneous enough to allow no probability to exceed 0.5. For the remaining samples, between 87.2 and 90.4 % were assigned to the correct tissue of origin. When a maximum likelihood criterion was used (Supplemental Table 5i), an Undetermined assignment was not allowed and between 85.7 and 87.9 % of all SAMMs were assigned to the correct tissue of origin. The breakdown by Groups is shown in Supplemental Tables 5j and 5k, when the probability of membership had to be at least 0.5 and using the maximum likelihood criterion, respectively. A comparison of Supplemental Tables 5b and 5j shows that with the exception of prostate samples, reducing the number of Undetermined

classification reduces the overall accuracy of the prediction. In the breast tumor samples, for example, the number of Undetermined samples decreases from 31 to 4. Four of these 27 samples (14.8 %) were assigned to the correct tissue of origin, and 23 were not. Since breast samples only represent 10.5 % of all samples considered in this study, the four samples moved to this category is larger than the number expected by chance (2.84), but this reduced the accuracy of prediction from 71.9 to 54.9 %.

We believe that integrating our approach with current diagnostic tools could serve to improve tumor classification scores. This would necessitate increasing the number of cancer genomes for each tissue type and including more tissue types in the analysis, but this analysis nevertheless constitutes a promising start. In addition, refinement of the SAMMs representing specific mutation mechanisms and increasing the number of MTMMs might also improve the tissue of origin classification, since maximizing the amount of contributory information should strengthen the prediction.

Conclusions

There are several aspects to this investigation that we believe set our analysis of cancer genome mutational spectra apart from all other reports to date. First, by analyzing pentamer motifs, we have captured the influence of two nucleotides upstream and downstream of the mutation sites without loss of discriminatory power. This information is stored as a 32 × 12 somatic autosomal mutation matrix (SAMM). Second, we constructed canonical MTMMs representing four of the most common mutational mechanisms, viz. oxidative damage, UV-induced CPD formation, methylation-mediated deamination and APOBEC-induced deamination. For the oxidative damage MTMM, we applied quantum chemical calculations to derive vertical ionization potentials, which were then used to establish mutational patterns for all relevant trinucleotide motifs. Of the 15 sample SAMMs that contained the strongest signature from this mechanism, 13 represented lung cancers. This suggests that loss of an electron is the rate-limiting step in this mutational mechanism. We believe this is a relevant conclusion, since it applies the findings accumulated over the past 30 years from the studies of electron transfer reactions on short DNA oligomers to the field of human cancer biology. Thus, the MTMM shown in Supplemental Figure 1a and Supplemental Table 1a may represent the actual pattern for oxidative damage. Third, by calculating the Manhattan distance between SAMMs from different samples, we constructed a Sammon map that provides an 'anatomical' representation of how cancer tissues are related to each other, and shows sub-clusters containing specific cancer types. Note that this is very different from measuring projections of a sample's SAMM along mechanistic or any other non-orthogonal signatures, since this procedure will not preserve distance. We believe that preserving the inter-sample distance is critical for achieving high-resolution clustering.

We show that cancer tissue preference exists for each MTMM (lung for oxidative damage, liver for photo-damage, pancreatic cancer for ^{5m}CpG deamination and breast cancer for APOBEC activity). Most importantly, Manhattan distances were able to achieve 89.5 % accuracy in placing tumor SAMMs of the same tissue type as nearest neighbors, and 92.2 % accuracy for the 810 SAMMs with a definitive classification from a distance-dependent 6-nearest neighbor algorithm, implying that our in-depth analysis of single-base substitution patterns nears the diagnostic power currently attained by clinical and pathological analyses. Thus, our approach may eventually augment current diagnostic procedures by helping to improve tumor classification scores. Nevertheless, the existence of prominent mutational signatures over and above the four mutational processes considered here and the finding that specific types of tumor were consistently misclassified highlight the need to further address the mechanisms underlying the origin of mutations in a tissue- or cell-type-specific fashion.