A substantial proportion of cancer cases present with a metastatic tumor and require further testing to determine the primary site; many of these are never fully diagnosed and remain cancer of unknown primary origin (CUP). It has been previously demonstrated that the somatic point mutations detected in a tumor can be used to identify its site of origin with limited accuracy. We hypothesized that higher accuracy could be achieved by a classification algorithm based on the following feature sets: 1) the number of nonsynonymous point mutations in a set of 232 specific cancer-associated genes, 2) frequencies of the 96 classes of single-nucleotide substitution determined by the flanking bases, and 3) copy number profiles, if available. We used publicly available somatic mutation data from the COSMIC database to train random forest classifiers to distinguish among those tissues of origin for which sufficient data was available. We selected feature sets using cross-validation and then derived two final classifiers (with or without copy number profiles) using 80% of the available tumors. We evaluated the accuracy using the remaining 20%. For further validation, we assessed accuracy of the without-copy-number classifier on three independent data sets: 1669 newly available public tumors of various types, a cohort of 91 breast metastases, and a set of 24 specimens from 9 lung cancer patients subjected to multiregion sequencing. The cross-validation accuracy was highest when all three types of information were used. On the left-out COSMIC data not used for training, we achieved a classification accuracy of 85% across 6 primary sites (with copy numbers), and 69% across 10 primary sites (without copy numbers). Importantly, a derived confidence score could distinguish tumors that could be identified with 95% accuracy (32%/75% of tumors with/without copy numbers) from those that were less certain. Accuracy in the independent data sets was 46%, 53% and 89% respectively, similar to the accuracy expected from the training data. Identification of primary site from point mutation and/or copy number data may be accurate enough to aid clinical diagnosis of cancers of unknown primary origin.

General information
State: Published
Organisations: Department of Systems Biology, Center for Biological Sequence Analysis, Cancer Systems Biology, INSERM, University College London
Number of pages: 13
Publication date: 2015
Peer-reviewed: Yes

Publication information
Journal: BMC Medical Genomics
Volume: 8
Issue number: 58
ISSN (Print): 1755-8794
Ratings:
Web of Science (2018): Indexed yes
Scopus rating (2017): CiteScore 3.41 SJR 1.688 SNIP 0.847
Web of Science (2017): Impact factor 3.317
Web of Science (2017): Indexed yes
Scopus rating (2016): CiteScore 2.96 SJR 1.661 SNIP 0.754
Web of Science (2016): Impact factor 2.848
Scopus rating (2015): CiteScore 3.4 SJR 1.638 SNIP 0.857
Web of Science (2015): Impact factor 2.726
Web of Science (2015): Indexed yes
Scopus rating (2014): CiteScore 3.48 SJR 1.911 SNIP 0.96
Web of Science (2014): Impact factor 2.873
Web of Science (2014): Indexed yes
Scopus rating (2013): CiteScore 4.26 SJR 2.213 SNIP 1.087
Web of Science (2013): Impact factor 3.914
Scopus rating (2012): CiteScore 4 SJR 1.937 SNIP 0.969
Web of Science (2012): Impact factor 3.466
Scopus rating (2011): CiteScore 4.08 SJR 2.129 SNIP 1.048
Web of Science (2011): Impact factor 3.693
Scopus rating (2010): SJR 2.166 SNIP 0.845
Web of Science (2010): Impact factor 3.766
Original language: English
Keywords: Cancer of unknown primary, Cancer genomics, Mutations