Population assignment from cancer genome profiling data

Qingyao Huang[1,2] **and Michael Baudis**[1,2]✉

[1]Institute of Molecular Life Science, University of Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland
[2]SIB, Swiss Institute of Bioinformatics, Winterthurerstrasse 190, 8057, Zurich, Switzerland

For a variety of human malignancies, incidence, treatment efficacy and overall prognosis show considerable variation between different populations and ethnic groups. Disentangling the effects related to particular population backgrounds can help in both understanding cancer biology and in tailoring therapeutic interventions. Because self-reported or inferred patient data can be incomplete or misleading due to migration and genomic admixture, a data-driven ancestry estimation should be preferred. While tools to map and utilize ancestry information from healthy individuals have been introduced, a population assignment based on genotyping data from somatic variation profiling of cancer samples is still missing.

We analyzed sequencing-based variation data from the 1000 Genomes project, containing 2504 individuals out of 5 continental groups. This reference was then used to extract population-biased SNPs used in genotyping array platforms of varying resolutions. We found that despite widespread and extensive somatic mutations of cancer profiling data, more than 90% of cancer samples can be correctly mapped to one of the population group when compared to their paired unmutated normals. Prefiltering samples for admixed individuals increased the accuracy to 96%.

This work provides a data-driven approach to estimate the population background from cancer genome profiling data. This proof-of-concept study will facilitate efforts to understand the interplay between population and ethnicity related genetic background and differences in understanding statistical and molecular differences in cancer entities with respect to possible hereditary contributions. The docker version of the tool is provided through "baudisgroup/tum2pop" in DockerHub and deposited in "baudisgroup/tum2pop-mapping" in GitHub.

Cancer genomics | Population background | SNP array
Correspondence: ***mbaudis@imls.uzh.ch***

Introduction

Cancer arises from the accumulation of genomic aberrations in dividing cells of virtually all types of tissues (somatic variations). The irregular cellular expansion and other hallmarks of cancer (1) can result from a plethora of mechanisms affecting multiple cellular processes. Some of the individual oncogenetic pathways can be initiated by exogenous factors, e.g. tobacco smoke or ultraviolet radiation (2). However, exposure to these carcinogenic factors contribute differently for people with different genetic background, which suggests that somatic variations can be influenced by inherited ("germline") variations (3, 4).

Statistics on cancer report considerable variation in incidence and prognosis between ethnicity groups (5–8). While such differences have been attributed to unequal social and economical circumstances influencing risk factors and therapeutic interventions, they may also reflect the impact of population specific genomic variants with predisposing effects on malignant transformation and phenotypic behaviour. Due to the late onset of most cancers, even high-penetrance Mendelian-type variants may not be purged by natural selection and can accumulate in particular populations. Such variants may play key roles in cancer development (9). Notably, mutations on BRCA1/2 genes confer a high risk to develop breast and ovarian carcinomas. Three founder mutations in Ashkenazi Jewish population cause the BRCA1/2 mutation prevalence to be 10-fold higher than all sporadic mutations in the general population (10, 11). Mitochondrial aldehyde dehydrogenase (ALDH2) encodes an enzyme in alcohol metabolism. Its "oriental" variant with 36% prevalence in East Asians, ALDH2*504Lys, increases risk for alcohol-related liver, colorectal and esophageal cancer by alcohol consumption (12, 13).

Many other studies have reported prevalent genetic variants in specific population groups which may contribute to the "racial" disparities in occurrence and prognosis (14–16). Other than these monogenic determinants, polygenic variation models for breast cancer which estimate the combined effect of multiple loci to be highly discriminatory in risk assessment, suggest the benefits of exploring genome-wide risk profiles (17). The potential impact of understanding the germline background of cancer genomes has also been demonstrated in a study which identified disease-associated chromosomal regions from only seven individual samples by using genome-wide relatedness/linkage mapping (18). This type of studies can be conducted population-wise, with sufficient number of samples from the same population/ethnicity group.

With the increasing number of available genome profiles and the decreasing cost to genotype clinical samples, the stratification between patients' genetic backgrounds has become feasible with the promise to guide therapeutic strategies and improve the clinical prognoses. Since several studies have demonstrated the relevance of considering an individual's genomic origin for preventive screening (reviewed in Foulkes *et al.* 2015 (11)), information about the population background of cancer patients may be an additional factor for individual therapeutic decisions as well as for the stratification of clinical study cohorts. A meta-analysis addressing the interplay between genetic background, cancer development and ther-

apeutic responses is desirable, not only for robust statistical associations in molecular target identification, but also for the rational design of studies incorporating informative biosamples.

The "population group" of a sample can be determined based on a geographical location associated with the sample; from self-reported "race", as commonly used in U.S. census data, or based on a computational estimate of ancestry by modelling population related genomic variants.

In the context of anonymized or pseudonymized research data, an approximation of a biosample's geographic origin can be achieved by using the location of the study's research facility or alternatively the contact address of its main authors. However, while these data can be easily retrieved, they may not provide an accurate representation of patients' origins for the purpose of population-specific ancestry mapping. Self-reported data is often inconsistent across studies, vague in category description (e.g. *"white"*, *"black"* v.s. *"Caucasian"*, *"African"*) and misleading when patients do not know the migration and admixture histories of their ancestors. Overall, when associating oncogenic molecular signatures with germline variations, information from the above sources lacks in relevant detail and consistency.

A better approach to population assessment would be the direct inference from genomic data. This has been shown previously for germline profiles, achieving 90% accuracy by using as few as 100 population-diverging single nucleotide polymorphisms (SNPs) (19), and nowadays is the standard methodology behind a number of commercial "ancestry" services. We hypothesise that a similar strategy can be applied to cancer genome data, despite the additional cancer-related somatic mutations which leads to both information loss (e.g. large scale homozygous or allelic deletions) and added noise (e.g. somatic mutations masking germline variants). An example of a cancer genome containing copy number loss and copy-neutral loss of heterozygocity (CN-LOH) events and its paired normal sample is shown in Figure 1. Additionally to a general test of feasibility, we also set out to benchmark population mapping procedures for heterogeneous datasets from different genotyping platforms, with varying SNP content.

Results and Discussion

We retrieved the genomic reference data from the 1000 Genomes Project (20), containing sequencing data of 2,504 individuals from 26 populations of five continental ancestries. SNPs of the selected array platforms were extracted from the sequencing data for reference samples. In order to achieve between-study consistency for selection of informative SNPs, we used a model-based approach (21) where an admixture model is optimized with the reference set for each genotyping platform. The allele frequency and ancestry fraction parameters were projected to the incoming cancer dataset of the same platform. Applying a random forest classification, we assigned the population label to the highest voted group and produced a score for the difference between highest and second highest percentage votes (Figure 2). We benchmarked our method with various normal and cancer datasets to demonstrate the feasibility and reliability of this approach.

Cross-platform benchmarking. We first used the original data from 1000 Genomes to validate the level of resolution needed for accurate population assignment from the pipeline. The number of SNPs per platform ranged from 10,204 (Affymetrix Mapping10K) to 934,946 SNPs (Affymetrix Genome Wide SNP 6). For all nine genotyping platforms (of seven levels of resolution), the model performed equally well in capturing the informative SNPs and predicting the population category. The 26 population groups are assigned into the 5 continental categories with low margin error for all genotyping platforms (less than 12 in 2,504 individuals) (Figure 3). When evaluating the 12 mis-classified individuals by cross-validation, we discovered that they were repetitively assigned into the same aberrant category; therefore, they were removed for the final implementation (Additional File 1). In addition, we removed 396 individuals from the random forest assignment based on the admixture background.

Benchmarking normal genome profile assignment with HapMap data. To validate the general ability to map population origins from non-cancer SNP datasets, we used 112 samples found in Gene Expression Omnibus (GEO (22)) belonging to the HapMap project (23) but not included in the reference set. While most assigned labels matched the metadata from HapMap, five samples labeled "European" were assigned to the "American" category (Table 1).

Table 1. Comparison of HapMap metadata and predicted population group.

	CEU	CHB	YRI
AFR	0	0	45
AMR	5	0	0
EAS	0	6	0
EUR	56	0	0

Columns indicate HapMap population labels. CHB for Han Chinese in Beijing, China. YRI for Yoruba in Ibadan, Nigeria

Paired cancer-normal comparison. The emphasis of the pipeline lies in the determination of population origin from cancer genome profiles carrying varying somatic mutations. Since the non-cancer samples could be correctly assigned according to HapMap categories, we validated the cancer genome based assignments in samples where normal genome profiles of the same patients (e.g. from peripheral blood or non-cancer tissue samples) were available as reference.

GEO data. From the GEO repository, we selected paired normal and cancer samples from 1219 individuals and compared the outcome of the population assignment. When including all individuals, 92.5% of the normal samples matched with paired tumor samples. After setting a threshold of normal samples with score > 0.2, 96.2% accuracy can be achieved for the remaining 762 individuals. When also setting the

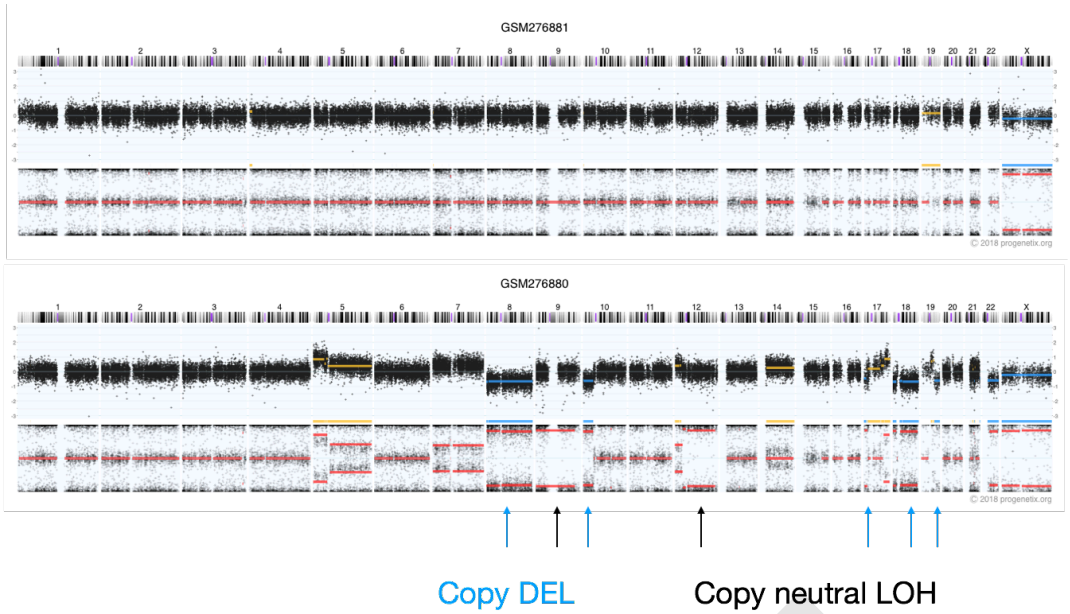

Fig. 1. CNV examples for a pair of normal/cancer samples. Compared to the normal sample (upper panel), the cancer sample (lower) has copy number loss in chr8, 10p, 18, 19qter, 22q and copy-neutral loss of heterozygocity in chr9,12q.

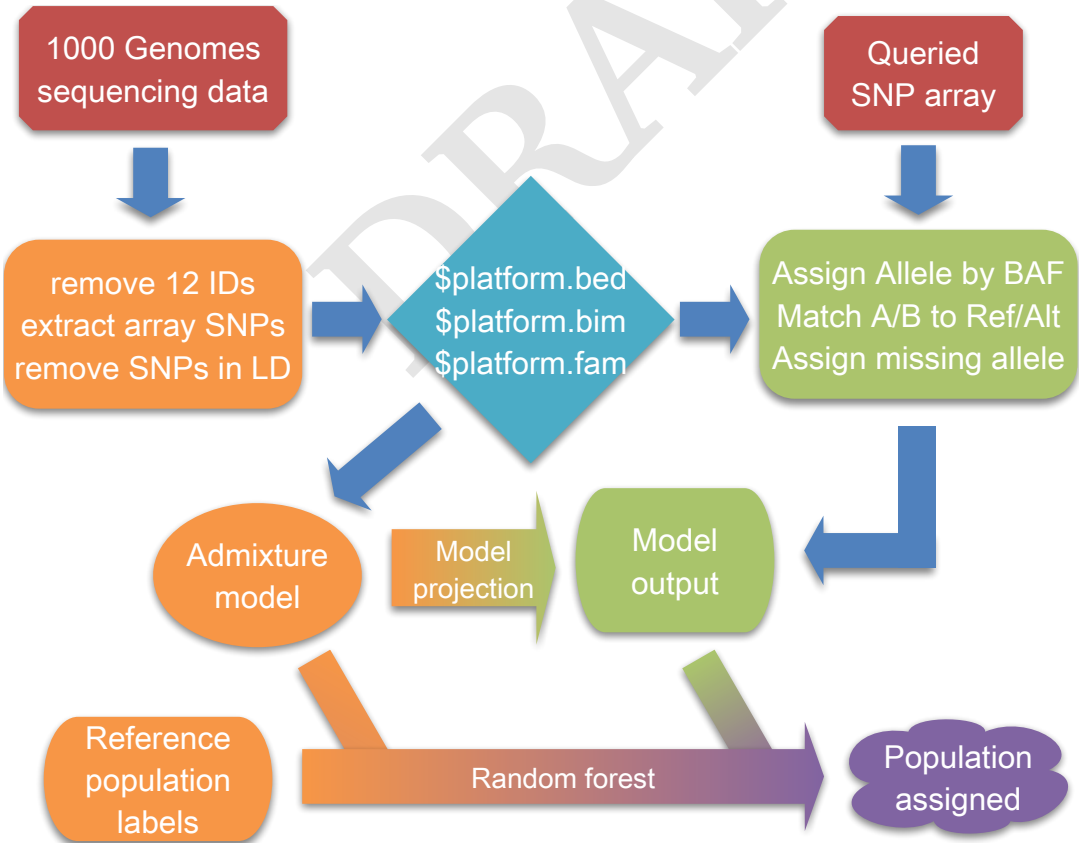

Fig. 2. Pipeline to derive population assignment for individual cancer samples.

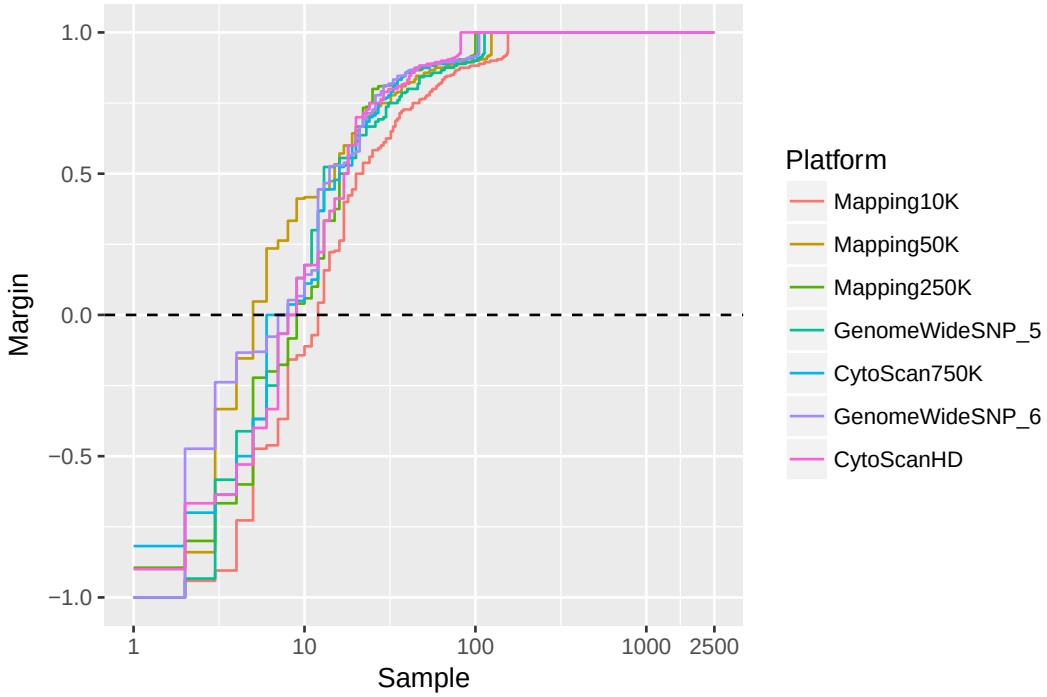

Fig. 3. Margin plot of prediction on platforms of seven different resolution.
Margin is defined as the difference between highest vote and the correct vote in random forest, a positive value indicates correct prediction.

score threshold for cancer samples to > 0.2, 99.0% of the now 721 remaining samples could be matched correctly (Figure 4). This comparison suggests that a correct assignment of cancer samples to a population category can be achieved, and that the level of accuracy increases with a lower admixture background of the individual.

TCGA data. We performed a similar measurement with 436 randomly selected individuals from the TCGA project (24), where at least one normal and one tumor sample per individual were available. 433 out of 436 (99.1%) individuals had matched tumor/normal categories, with the three outliers switching from EUR to AMR between samples 5. Additionally, we compared our results with the values of the "race" attribute provided in the TCGA metadata. There, six categories are being distinguished: "American Indian or Alaska native", "Asian", "Black or African American", "Native Hawaiian or other Pacific Islander", "White or Not Reported". The relevant ratio of these groups is shown in Figure 5. Most assignments were accurate: *EUR* samples were mostly "White"; *AMR* has mostly "White" and a few "African American" samples; *SAS* are all "Asian" samples. Additionally to "Asian" samples, "American Indian or Alaska Natives" or "Pacific Islanders" were all assigned to *EAS*.

Together, these two validation tests confirmed that a high assignment accuracy for cancer samples can be achieved, mirroring the assignment of their corresponding germline samples. Since samples with scores lower than the threshold had a highly admixed background, we also noted a mixture between AMR and EUR observed in both normal samples and cancer samples. We attribute this to the complex, recent admixture events in the last hundreds of years in the locations from where individuals were recruited as AMR in reference data (e.g. Colombia, Puerto Rico).

Self-reported ethnicity metadata. After the validation of the method, we tested the accuracy of self-reported metadata from various sources deposited in GEO. We retrieved a total of 1724 samples with interpretable self-reported metadata. Out of those, 1523 samples (88.3%) were correctly assigned. When setting a threshold of the assignment score to 0.2, 92.7% (1310 out of 1412) samples were correctly assigned (Figure 6). This increase on matched assignments is more limited compared to the previous validation tests on GEO data, suggesting the contribution of curation errors and inaccuracies in self-reported ethnicity metadata.

Conclusions

We demonstrate the feasibility and accuracy of assigning population group provenance based on SNP genotyping array data of cancer samples, where somatic mutations obfuscate parts of the ancestry related SNP signal. This work can facilitate meta-analysis of available cancer data with respect to the association of the genetic background to cancer specific mutations or, as proxy, to the correct assignment of sample provenance. In addition, our method provides the basis for subsequent haplotype phasing and refinement of genomic landscape for emerging somatic variation. Concerning the delicate balance between data utilization and confidentiality protection, it has not escaped our notice that the relative feasibility of such an approach suggests the potential of individual

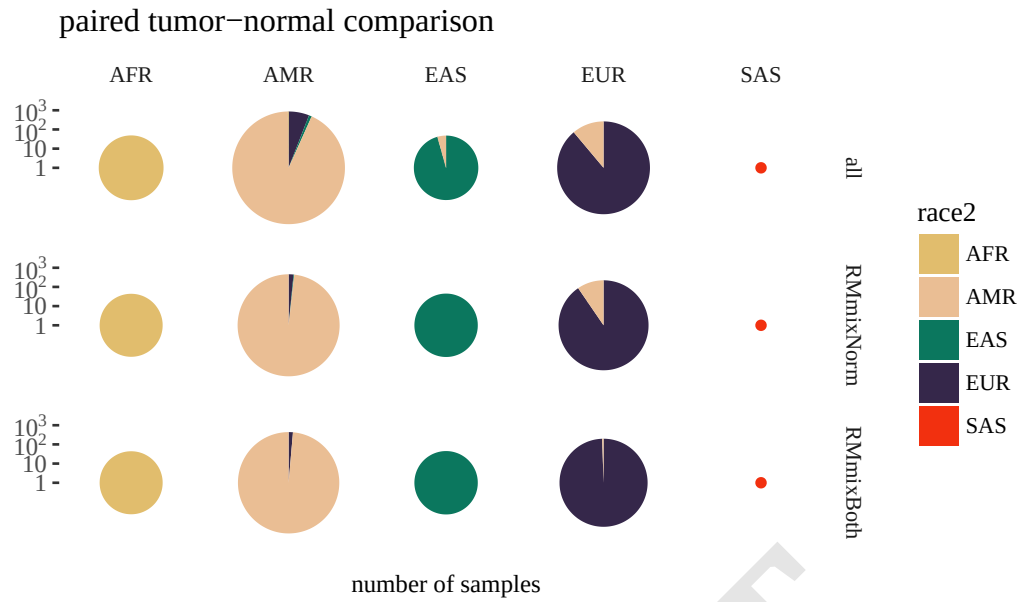

Fig. 4. Accuracy of assignment with paired tumor and normal samples from GEO.
1219 individuals from GEO with paired tumor and normal samples were examined. "all" indicates the total 1219 individuals. "RMmixNorm" indicates results for 762 individuals, after 447 individuals were removed because of low score in the normal sample. "RMmixBoth" indicates results for 721 individuals, after additional 41 individuals were removed due to low score found in the cancer sample.

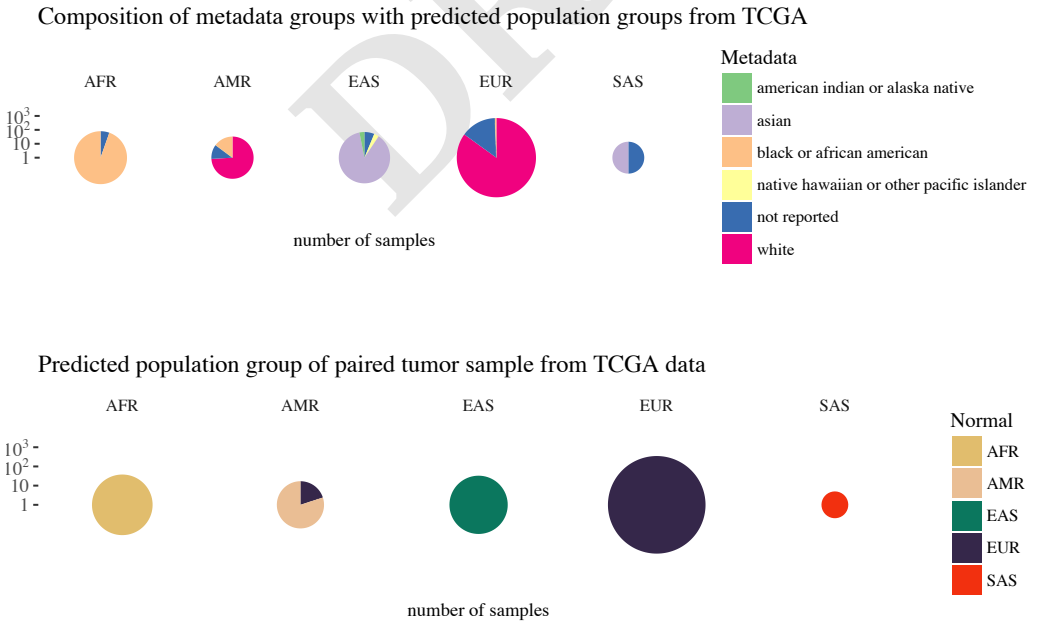

Fig. 5. Accuracy of assignment with paired tumor and normal samples and comparison to metadata from TCGA project.
436 individuals with paired tumor and normal samples from TCGA project were recruited. Upper panel compares the assignment with metadata and lower panel indicates the matching proportion of predicted population groups between tumor and normal samples of the same individuals.

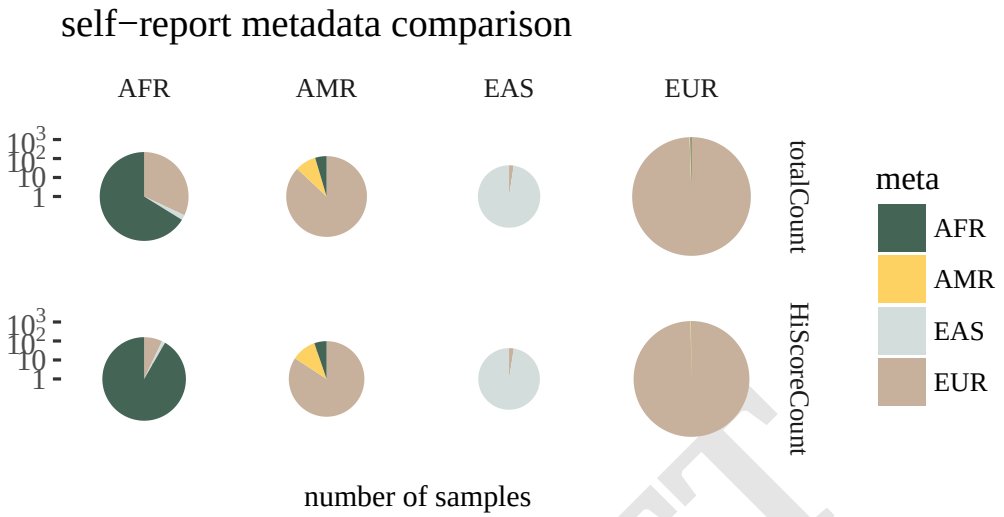

Fig. 6. Accuracy of assignment with self-report metadata.
883 samples from GEO which contain adequate population/ethnicity metadata were examined. The "totalCount" indicates all samples. The "HiScoreCount" indicates the samples which have score > 0.2.

re-identification from cancer genotyping data.

Methods

Data preparation. Reference sequencing data are provided by 1000 Genomes Project, a publicly available reference catalogue of human genotype variation. Data used for benchmarking are accessed through arrayMap database (25), using a collection of re-processed genotyping series from the GEO repository, and the TCGA data repository.

Reference data preparation. The SNP positions for each platform were acquired from Affymetrix annotation files. The allele information was extracted for all positions with vcftools. The 12 mislabeled or admixed individuals were removed from the reference dataset, leaving 2,492 individuals. The SNP positions with duplicated rsIDs in annotation files were removed. The reference/alternative alleles were flipped according to the SNP array annotation. Sites with minor allele frequency (MAF) of less than 5 percent were removed. SNPs were subsequently pruned with variance inflation factor(VAF) at 1.5 with a sliding window of 50bp and a 5bp shift of window at each step using PLINK 1.9. The result files were stored as PLINK output for each platform in .bed, .bim and .fam formats, of which the .bim files were used to extract SNP positions from target data.

Target data preparation. The SNP array data were processed with ACNE R package (26) to extract allele-specific copy numbers as B-allele frequencies (BAF). SNPs were labeled as homozygous A, heterozygous AB or homozygous B by the BAF value in ranges 0-0.15, 0.15-0.85 or 0.85-1, respectively, to allow both for noise and expected aneuploidy in the biosamples.

Admixture model. While many approaches use principle component analysis (PCA) to select informative SNPs for population assignment, deriving them prior to clustering methods, either by removing correlated SNPs (with Pearson's $r > 0.99$) or by global fixation index (Fst > 0.45), results in

varying remaining SNPs between datasets.

We used the allele information output from the reference panel to generate an admixture statistical model (21), which estimates the contribution of each SNP to the population category by alternately updating allele frequency and ancestry fraction parameters. Models were built with choosing the number of theoretical ancestor, K = 6, considering the cross-validation error, iteration steps and runtime. The ancestry fraction plot for reference individuals demonstrates a proper information extraction to distinguish the five continental categories. By projecting a correspondingly learned model derived from the reference dataset to a new sample with the corresponding platform, a robust and consistent output with 6 ancestry fractions was generated.

Random Forest label assignment. 396 samples were selected out of the random forest training set, due to the admixture structure found in the ancestry fraction (Additional File 2). 2,096 (2,492 less 396) samples were used as training set. The six ancestry fractions from the reference population per platform were used to build a random forest model to predict the five population categories. The model was then used to predict a label to the population category from 6 fractions of the target sample. The score was calculated as the difference in percentage votes between the best and the second best predicted labels.

Abbreviations

ALDH2: Aldehyde Dehydrogenase 2 Family (Mitochondrial) **AMR:** American **AFR:** African **BAF:** B allele frequency **EAS:** East Asian **EUR:** European **Fst:** Fixation Index **MAF:** Minor allele frequency **GEO:** Gene Expression Omnibus **PCA:** Principle Component Analysis **SAS:** South Asian **SNP:** Single nucleotide polymorphism **TCGA:** the Cancer Genome Atlas

Competing interests

The authors declare that they have no competing interests.

Funding

QH has been supported by University of Zurich with "Forschungskredit CanDoc" fellowship.

Author's contributions

QH developed the tool and performed benchmarking. MB conceived the project. Both authors contributed to data assembly and writing and editing of the manuscript.

ACKNOWLEDGEMENTS

We thank Paula Carrio Cordo for the metadata curation and Bo Gao for technical support and helpful discussions. We thank the Zurich Bioinformatics group and ELIXIR Population Genomics for valuable comments.

Additional Files

Additional file 1 — Removed 12 individuals. 12 individuals from the reference 1000 Genomes database, with admixed or mislabeled origin and therefore removed from the pipeline, leaving 2,492 in the reference.

Additional file 2 — Removed 396 individuals. The 396 individuals contain a admixture background which could mix up with another population category, including 225 AMR (101 Puerto Rican, 88 Colombian, 32 Mexican-American, 4 Peruvian), 40 SAS (30 Punjabi, 10 Gujarati), 33 AFR (9 African-Caribbean, 24 African-American SW).

Bibliography

1. D Hanahan and RA Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5): 646–674, 2011.
2. Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415, 2013.
3. HT Lynch and A de la Chapelle. Hereditary colorectal cancer. *N Engl J Med*, 348(10): 919–932, 2003.
4. J Zhang, KE Nichols, and JR Downing. Germline mutations in predisposition genes in pediatric cancer. *N Engl J Med*, 374(14):1391, 2016.
5. D Max Parkin, Paola Pisani, and J Ferlay. Global cancer statistics. *CA: a cancer journal for clinicians*, 49(1):33–64, 1999.
6. G Danaei, S Vander Hoorn, AD Lopez, CJ Murray, M Ezzati, and Risk Assessment collaborating group (Cancers Comparative. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet*, 366(9499):1784–1793, 2005.
7. RL Siegel, KD Miller, and A Jemal. Cancer statistics, 2017. *CA Cancer J Clin*, 67(1):7–30, 2017.
8. Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30, 2018.
9. Steven A. Frank. Genetic predisposition to cancer – insights from population genetics. *Nature Reviews Genetics*, 5:764 EP –, Oct 2004. Review Article.
10. Yoshio Miki, Jeff Swensen, Donna Shattuck-Eidens, P Andrew Futreal, Keith Harshman, Sean Tavtigian, Qingyun Liu, Charles Cochran, L Michelle Bennett, Wei Ding, et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266 (5182):66–71, 1994.
11. William D. Foulkes, Bartha Maria Knoppers, and Clare Turnbull. Population genetic testing for cancer susceptibility: founder mutations to genomes. *Nature Reviews Clinical Oncology*, 13:41 EP –, Oct 2015. Review Article.
12. Hui Li, Svetlana Borinskaya, Kimio Yoshimura, Nina Kal'ina, Andrey Marusin, Vadim A Stepanov, Zhendong Qin, Shagufta Khaliq, Mi-Young Lee, Yajun Yang, et al. Refined geographic distribution of the oriental aldh2* 504lys (nee 487lys) variant. *Annals of human genetics*, 73(3):335–345, 2009.
13. Philip J Brooks, Mary-Anne Enoch, David Goldman, Ting-Kai Li, and Akira Yokoyama. The alcohol flushing response: an unrecognized risk factor for esophageal cancer from alcohol consumption. *PLoS medicine*, 6(3):e1000050, 2009.
14. Tanya Keenan, Beverly Moy, Edmund A. Mroz, Kenneth Ross, Andrzej Niemierko, James W. Rocco, Steven Isakoff, Leif W. Ellisen, and Aditya Bardia. Comparison of the genomic landscape between primary breast cancer in african american versus white women and the association of racial differences with tumor recurrence. *Journal of Clinical Oncology*, 33 (31):3621–3627, 2015. doi: 10.1200/JCO.2015.62.2126. PMID: 26371147.
15. Jiaying Deng, Hu Chen, Daizhan Zhou, Junhua Zhang, Yun Chen, Qi Liu, Dashan Ai, Hanting Zhu, Li Chu, Wenjia Ren, Xiaofei Zhang, Yi Xia, Menghong Sun, Huiwen Zhang, Jun Li, Xinxin Peng, Liang Li, Leng Han, Hui Lin, Xiujun Cai, Jiaqing Xiang, Shufeng Chen, Yihua Sun, Yawei Zhang, Jie Zhang, Haiquan Chen, Shijian Zhang, Yi Zhao, Yun Liu, Han Liang, and Kuaile Zhao. Comparative genomic analysis of esophageal squamous cell carcinoma between asian and caucasian patient populations. *Nature Communications*, 8(1): 1533, 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-01730-x.
16. Wensheng Zhang, Andrea Edwards, Erik K. Flemington, and Kun Zhang. Racial disparities in patient survival and tumor mutation burden, and the association between tumor mutation burden and cancer incidence rate. *Scientific Reports*, 7(1):13639, 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-13091-y.
17. Paul D. P. Pharoah, Antonis Antoniou, Martin Bobrow, Ron L. Zimmern, Douglas F. Easton, and Bruce A. J. Ponder. Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genetics*, 31:33 EP –, Mar 2002. Article.
18. Anders Albrechtsen, Thorfinn Sand Korneliussen, Ida Moltke, Thomas van Overseem Hansen, Finn Cilius Nielsen, and Rasmus Nielsen. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic epidemiology*, 33(3):266–274, 2009.
19. Rust Turakulov and Simon Easteal. Number of snps loci needed to detect population structure. *Human heredity*, 55(1):37–45, 2003.
20. A Auton, LD Brooks, RM Durbin, EP Garrison, HM Kang, JO Korbel, JL Marchini, S McCarthy, GA McVean, GR Abecasis, and 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.
21. DH Alexander, J Novembre, and K Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655–1664, 2009.
22. R Edgar, M Domrachev, and AE Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, 2002.

23. International HapMap Consortium. The international hapmap project. *Nature*, 426(6968): 789–796, 2003.
24. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.
25. H Cai, N Kumar, and M Baudis. arraymap: a reference resource for genomic copy number imbalances in human malignancies. *PLoS One*, 7(5):e36944, 2012.
26. Maria Ortiz-Estevez, Henrik Bengtsson, and Angel Rubio. Acne: a summarization method to estimate allele-specific copy numbers for affymetrix snp arrays. *Bioinformatics*, 26(15): 1827–1833, 2010.