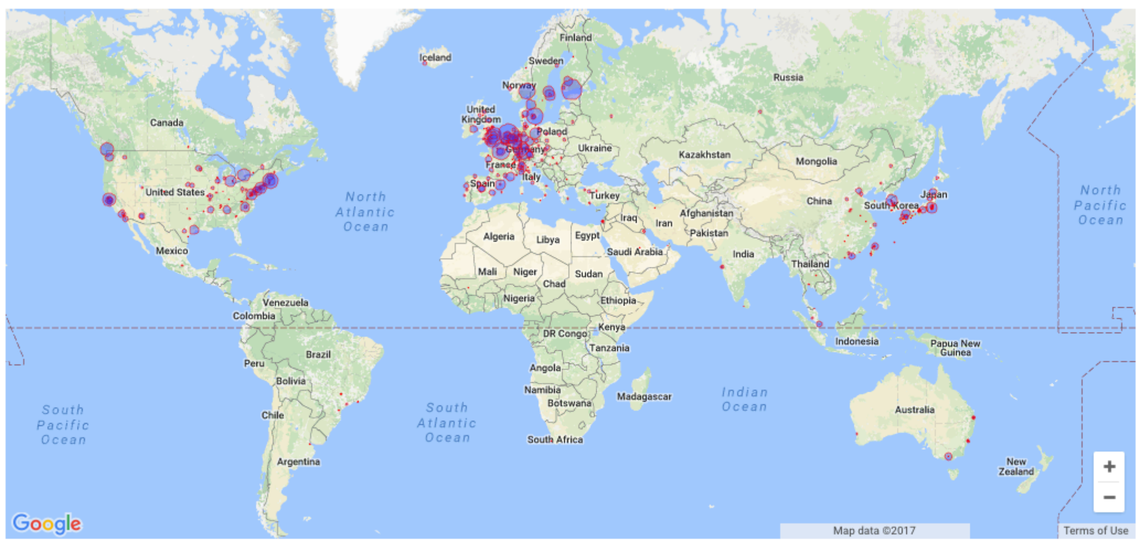
Integrative analysis of cancer genome profiling data to study the interplay of genetic background and molecular mechanisms in cancer

Qingyao Huang, Michael Baudis. Institute of Molecular Life Science, University of Zurich

Malignant neoplasias are based on the accumulation of mutations in cells during the lifetime of an individual (“somatic mutations”), which can be influenced by inherited (“germline”) genome variations. As tumor types and incidences differ among human populations, the genetic background of individuals could be one factor influencing somatic variation and subsequent tumorigenesis. In recent years, a large amount of cancer genome studies has been published, in thousands of tumor series analyzed by various genome screening techniques. However, most studies have been focused on individual tumor types and have been limited to genomic backgrounds of a few human populations. So far, the systematic analyses and integration of multiple available data sources are lacking. In this project, we perform a meta-analysis of the curated oncogenomic data from the arrayMap database, derived from various types of genomic arrays, and combine genomic profiles with epidemiological data to evaluate the population specificity of genome variations in cancer. From sequencing data of 26 populations world-wide from 1000Genome project, we extract the SNP markers corresponding to Affymetrix platforms and use them for subsequent sample analysis. First, we show that using admixture analysis, the population classification is accurate even from low-resolution arrays (10k markers). This will append genome-derived population information to the Progenetix database, as an addition layer to the geographic location of the publication-affiliated institute. As next step, we will link different types of chromosomal aberration (e.g. CN-LOH) to the identified population group to discover potential population-specific oncogenic patterns.

Geographical distribution of cancer research

Our group develops and maintains the Progenetix database, which curates and represents all publications, referring to cancer genome profiling experiments, including genomic arrays and whole genome/exome sequencing experiments. World-wide, cancer genomics studies conducted are far from evenly distributed in the geographic location, with hubs being North America, Europe and East Asia.



The map displays the geographic distribution (by corresponding author) of the 96703 genomic array, 36747 chromosomal CGH and 6274 whole genome/exome based cancer genome datasets. The numbers are derived from the 2993 I 2993 publications registered in the Progenetix database.

Figure 1: Worldwide distribution of cancer genomics data

Platform-wise benchmarking

With the hypothesis that ethnicity may play a role in wiring cancer development, we start by assigning population groups to samples. We perform admixture analysis on 2504 samples from phase1 of 1000Genome project, which are used as reference set (Fig. 2).

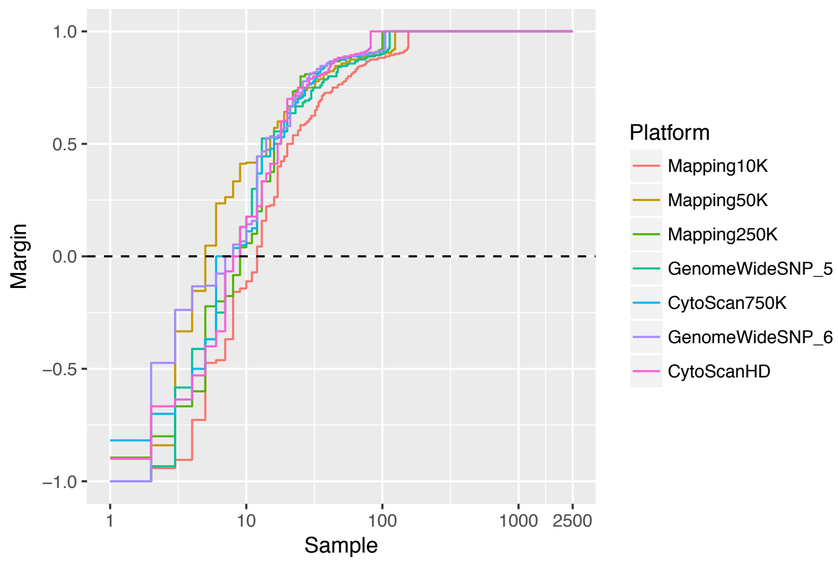


Figure 3: Evaluation of classification for samples from 1000Genome project. Margin is the proportion of votes for the correct class minus maximum votes for other classes. Positive margin means correct classification. For lowest-resolution platform (Mapping10K), the classification error rate is 0.44%

Extracting markers from various platforms from the sequencing data, we prove the accuracy of inferring ethnicity information from SNP arrays.

Challenges and Outlook

1. Assignment of samples with strong admixture background can prove difficult.
2. In mixed samples (subclonal cancer/normal), total and allelic copy number will diverge from integer. Algorithms are needed to robustly call patterns.
3. With the completion of population assignment, we will append accurate population information to the current Progenetix database, and
4. After evaluating the current copy number calling algorithms, we will develop algorithm which operates across platforms, tolerates subclonality, and identify population-specific genomic patterns and hotspot regions.

Population Admixture analysis

shows a clear separation between super-population, e.g. South Asia, Europe, East Asia, South America and Africa, and sample series has a mixed collection of population groups.

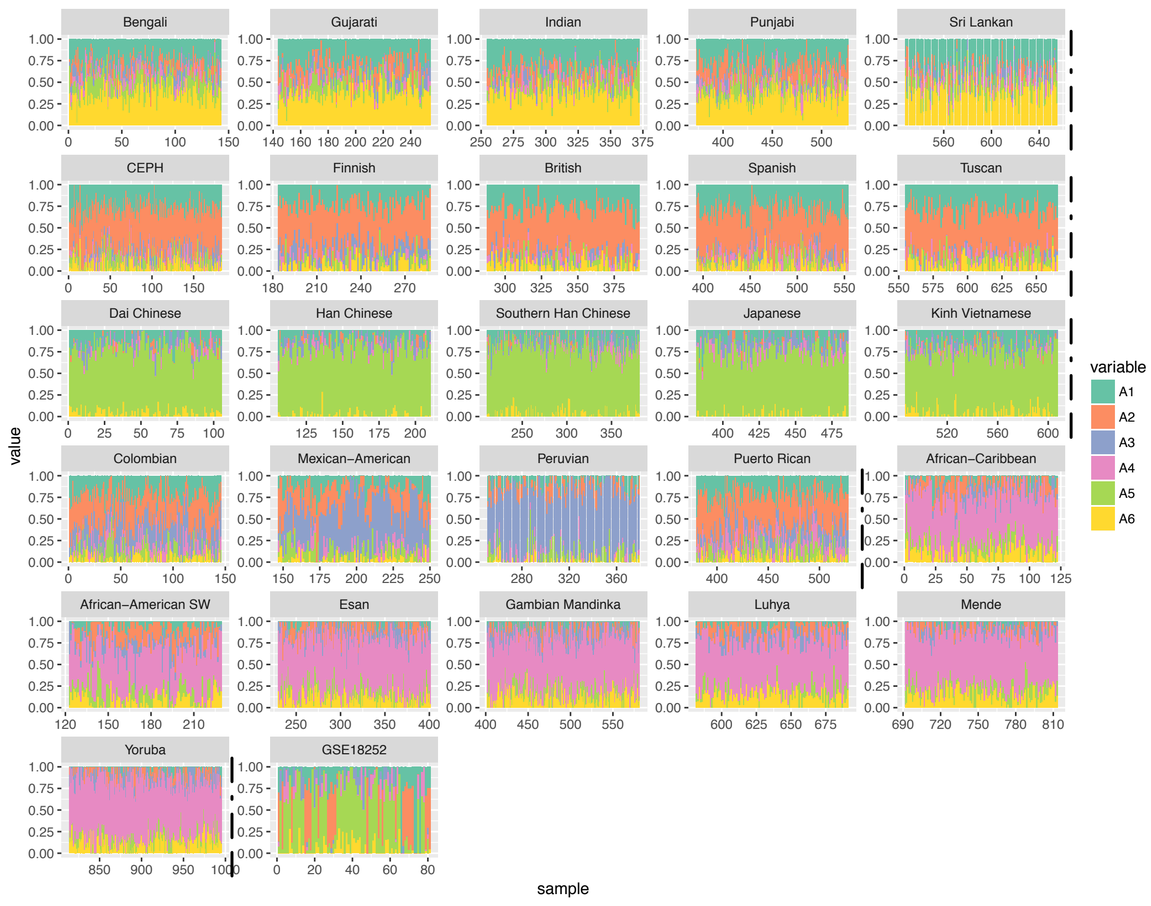


Figure 2: Admixture analysis of 2504 samples from 1000 Genome as reference of 26 subpopulations from 5 super populations (South Asia, Europe, East Asia, South America, Africa separated by dashed line), together with a GEO series GSE18252 derived from Genentech, South San Francisco. Each colour indicates a theoretical ancestral population. Analysis performed with A=6, with highest cross-validation score.

Genomic Pattern Analysis

Available in arrayMap database, we extract the total and allele-specific copy number (CN) from the available Affymetrix SNP platforms. Combining these two types of CN data, we can discern somatic variation patterns, e.g. total copy number gain/loss, copy neutral-loss of heterozygosity (CN-LOH), aka uniparental disomy (UPD), as well as mosaic loss/gain. With a collection of 70,002 tumor samples of various ICD-O types, we expect to summarise these patterns and in a population-specific manner.

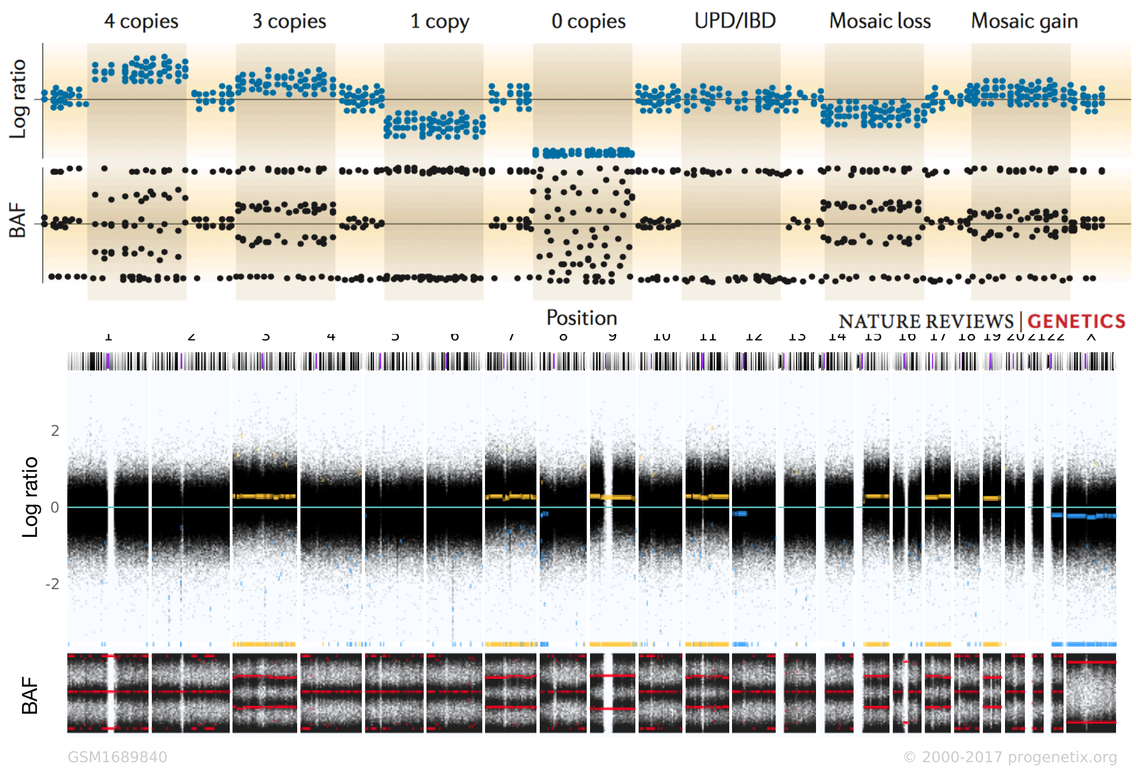


Figure 4: Total and allele-specific copy number, in terms of “Log ratio” of sample to reference copy number of base 2, and B-allele frequency (BAF). Upper: scenarios of chromosomal aberration, from Alkan et al., 2011. Lower: Data from Arraymap database, yellow for copy gain, blue for copy loss.

Universität Zürich[UZH]

Swiss Institute of Bioinformatics

life science zurich

progenetix

arrayMap

Reference:

1. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011
2. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009
3. Bengtsson H, Simpson K, Bullard J, et al. aroma.affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. Tech. rep. 745. Department of Statistics, University of California, Berkeley, Feb. 2008.