Trends in Cancer

CellPress

Science & Society

New York's Polyethnic-1000: a regional initiative to understand how diverse ancestries influence the risk, progression, and treatment of cancers

Nicolas Robine [1,*] and Harold Varmus [1,2,*]

Research consortia can help to repair deficiencies in knowledge about the influence of inherited genetic diversity on disease. The New York Genome Center (NYGC) recently established Polyethnic-1000 (P-1000), a multi-institutional collaboration to study hereditary factors affecting several types of cancer. Here, we describe its rationale, organization, development, current activities, and prospects.

Cancers are diseases of the genome, caused and influenced by genetic variation – both inherited variants and those acquired by somatic mutation – and by environmental and behavioral factors such as diet, exposure to mutagens, exercise, age, and reproductive strategies. Extensive efforts have been organized to identify the genetic variations that contribute to the initiation and progression of cancers, producing large data sets that have been especially useful for associating somatic mutations with several common neoplasms, improving diagnosis and therapy, and in the identification of some inherited alleles that confer significant risks of cancer [1–3]. However, most of the patients included in these studies are of Western European descent, reducing opportunities to study cancer in the context of the broad expanse of human diversity. Although there are a few notable exceptions [4–6], most have given inadequate attention to inherited, as opposed to acquired, genomic variation. Furthermore, despite some recent exceptions [7,8], most of the DNA sequencing has been limited to the protein-coding domains (exomes) of the genome or to panels of selected genes, ignoring vast regions of the genome.

To begin to address these deficiencies, the NYGC has assembled many of this large metropolitan region's hospitals and academic health centers to form a research consortium and to take advantage of both the rich ethnic diversity of the area and the NYGC's proficiency in whole-genome sequencing (WGS) and computational biology. Here, we summarize the assembly of this collaborative program, called P-1000, report findings from a small pilot program to demonstrate functionality, and describe the development of a peer-reviewed research program that has begun to study the contributions of ancestry to the biology and control of several types of cancer.

Building the P-1000 Consortium

The NYGC was founded in 2011 by 12 academic institutions to establish a center of excellence in genomics and bioinformatics in the New York region. In 2016, the NYGC established the Genome Center Cancer Group (GCCG) to encourage representatives of the now 19 academic partners to initiate large collaborative projects in cancer genomics. The GCCG has launched two initiatives to harness the strength of the center's partner institutions and the large, diverse population of the metropolitan area: P-1000, which is described in detail here, and the Very Rare Cancer Consortium (VRCC), which aspires to study the genetic factors that affect the initiation and progression of cancers that are diagnosed in less than one person per million per year[i].

To explore the feasibility of conducting a study of associations between inherited genetic variants and features of several neoplasms with a multi-institutional consortium in a large urban area, we established a P-1000 Steering Committee and several working groups comprising motivated physicians and scientists from GCCG-member institutions and the NYGC. Names of the participants can be found in the Acknowledgments section in the supplemental information online. These groups sought to build a framework for the conduct of the P-1000 initiative.

An important initial objective was the recruitment into the consortium of hospitals that were not previously affiliated with the NYGC; these healthcare facilities were mainly situated outside of the borough of Manhattan, served many patients from ethnic-minority communities, and provided their staff with limited time and resources for research, yet often belonged to networks established by academic health centers (Figure 1 and Table S1 in the supplemental information online). To ensure the appropriate provision of tumor samples from the participating hospitals, a clinical protocol was developed by a P-1000 working group and approved for multiple institutions by an institutional review board (IRB) organized by the Biomedical Research Alliance of New York (BRANY). In addition, a project-dedicated, part-time pathologist was recruited from a participating hospital at SUNY Downstate Medical Center to certify the clinical diagnosis of all samples to be studied in the early phase of the research program. In addition, standardized protocols were established for the procurement and shipping of clinical materials, for DNA sequencing and RNA sequencing, and for the assembly, storage, access, and use of genomic data derived from patient samples.

Testing the function of the consortium

To assess the functionality of this system, an early trial phase of P-1000 was launched to

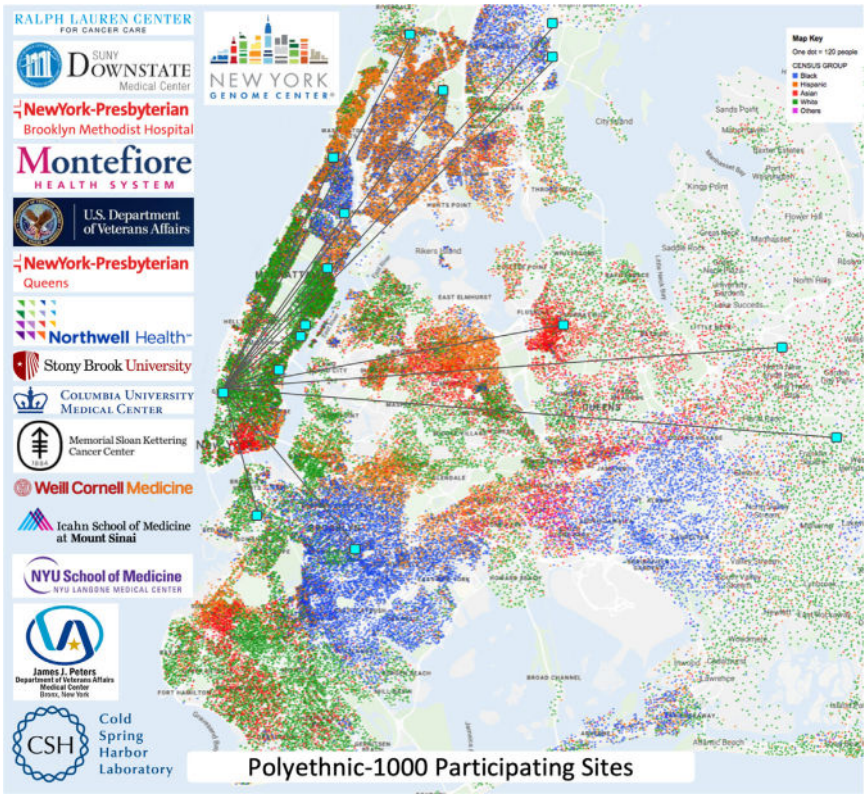

Figure 1. The institutions that participate in Polyethnic-1000, displayed on a map representing the self-declared racial and ethnic diversity of New York City according to the 2010 census. Source is *The New York Times*[ii].

obtain and study at least 100 stored tumor samples from the participating hospitals. Over the course of several months, 176 archival tumor samples, formalin fixed and paraffin embedded, from patients self-identified as other than 'white', were sent to the NYGC from participating hospitals (Table S1 in the supplemental information online) for whole-exome sequencing and RNA-seq. These tumors included 39 cancer types from 19 tissue sites, and the findings were analyzed using the NYGC pipeline [9] for mutations presumed to be somatic. For simplicity of design, normal tissues were not included in this preliminary exercise.

As expected, many single-nucleotide changes, short insertion and deletion mutations, gene fusions, and copy-number variations were detected in these samples, often affecting genes known to be altered in human cancers. For example, five tumors were classified including four colon adenocarcinoma and one uterine endometrioid carcinoma – three from African-American patients, one from a Hispanic patient, and one from an 'admixed American' patient – as microsatellite instable [10]. Using MSKCC's OncoKB annotations [11], we detected 'Tier 1' variants that define eligibility for FDA-approved drugs in nine samples (affecting the *EGFR*, *PIK3CA*, *BRAF*, and *FGFR3* genes), 'Tier 2' variants in five samples, and 'Tier 3' variants in 49 samples.

This pilot study was primarily a test of our ability to obtain and analyze tissue samples from the several hospitals in the P-1000 Consortium, but it was not designed to reveal new associations of genetic variations, inherited or somatic, with cancer biology. The findings, however, did allow us to demonstrate the power of genome sequencing to document ancestry in a way that is more informative than the common practice of self-identification with a single broad category of race, such as 'white', 'Asian', or 'African-American'.

The ethnic diversity of New York City is dramatically displayed by the image in Figure 1, which is based on self-declared categories of race, ethnic, and geographical origins commonly used in the USA. While this kaleidoscope of races and origins implies that the region's patient population is suitably diverse for the kinds of studies envisioned by the GCCG, the self-assignments are neither sufficiently reliable nor sufficiently nuanced to provide the information about ancestry necessary for a modern study of hereditary factors that might affect the initiation, progression, or treatment of malignancies.

To make a more rigorous assessment of genetic ancestries, we applied the ADMIXTURE software [12], using the reference populations from the 1000 Genomes Project [13], directly on the tumor samples used in our pilot project. This approach delivers an estimate of percentage of ancestry at the continental level (Figure 2, middle panel) and subregions (lower panel). In the top panel of Figure 2, we show self-described race or ethnicity for the corresponding patients to compare with our assessment of 'genetic ancestry'. We observe that genetic information nearly always correlates at least partially with self-identified origins, but that the genetic classifications are more specific and allow the identification of mixed ancestry. The mixtures imply matings in recent genealogical history between forebearers with distinct ancestries belonging to different populations. An understanding of the functions and origins of relevant genomic components will be required to establish

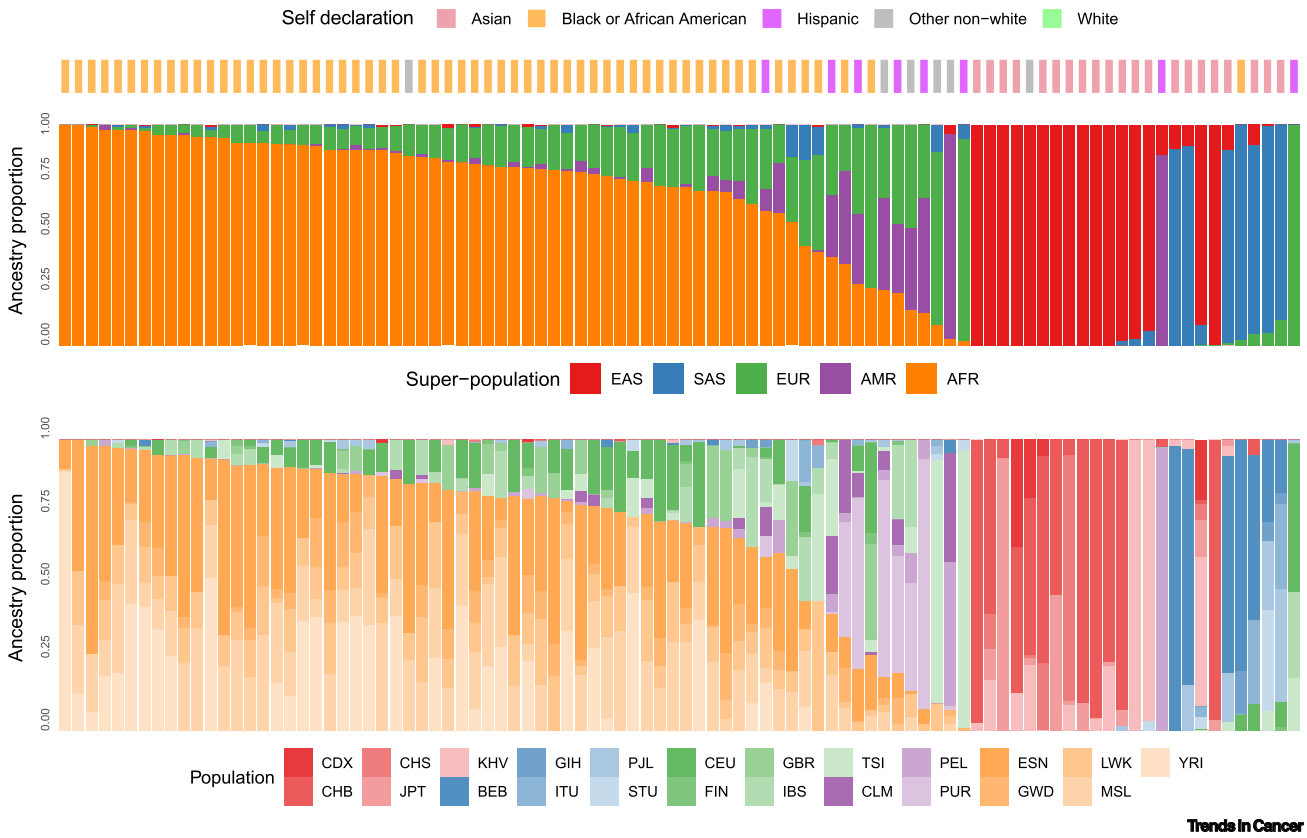

Figure 2. Results of genetic ancestry estimation on the Polyethnic-1000 (P-1000) pilot study. Top row represents self-declared race. Middle row represents 'continental-level estimates' (red, East Asians; blue, South Asians; green, Europeans; purple, Admixed American; orange, Africans). Bottom row represents 'population-level estimates' (using the 1000 Genomes three-letter code for the reference populations[iii]).

associations between ancestry and the features of a particular type of cancer.

Initiating research by the consortium

Having formed the consortium and shown that we can process samples from its member institutions for genomic analysis, we launched a research program with grants of modest size and duration. The availability of such awards – ranging from US$200 000 to US$500 000 for 2 years – allowed investigators in the consortium to compete for support of new or recently initiated projects designed to identify ancestry-associated genetic determinants in specific types of cancers. The level of interest among members of the consortium was reflected in the receipt of 21 letters of intent, followed by 13 full applications from mostly multi-institutional groups of investigators in response to a request for applications. Since our financial support was limited initially to a few philanthropic contributions, supplemented later by donations from some of the participating institutions, we were unable to award funds to all the applicants. We therefore recruited a panel of experienced investigators, all of whom work outside New York at institutions independent of the consortium, to review the applications and to recommend priority rankings. We funded seven of the competing groups, and they have now commenced work on at least eight different types of cancer (Table S2 in the supplemental information online). Most of the studies are focused on African-American patients; one addresses lung cancer in East Asian patients.

One of the objectives of P-1000 is to expand the use of the sophisticated methods developed at the NYGC for the determination and analysis of WGS of DNA from normal and tumor tissues. With support from Illumina, Inc., we plan to perform WGS on about 1000 tumor/normal pairs of tissues. The resulting data set would be among the largest and most diverse collection of full-genome pairs available in oncology. It will be housed initially at the NYGC, with access for all members of the P-1000 Consortium and then for the entire research community.

Research enterprises of the type illustrated by P-1000 require an unusual level of coordination, oversight, financial support, and regulatory adherence, advice from multiple disciplines of biology, medicine,

and social science, and a cooperative spirit among investigators in the consortium. The P-1000 Steering Committee has tried to promote these attributes in several ways. Our senior cancer biologists serve as advisors for the principal investigators on all grants. We have recruited two faculty members from our academic institutions to serve as part-time 'Cancer and Ethnicity Scholars', helping to solve problems encountered by the seven research groups. We hold bimonthly meetings (currently only online) for all investigators in the P-1000 Consortium to share information, to build collaborative ties, to hear lectures on related topics by investigators working in other locations, and to discuss common difficulties encountered in P-1000 projects. We have consulted social scientists to seek advice about better ways to engage with patients and their advocates. Also, we continue to seek the additional financial support that will be required to bring such work to the point at which medically significant conclusions can be reached.

Prospects for P-1000

We are conscious of the magnitude of the efforts required to gather enough genetic and clinical information to permit conclusions that will deepen our understanding of oncogenic mechanisms or promote changes in clinical practices such as prevention, risk assessment, diagnosis, or treatment. It is unlikely that studies of even the large and diverse population of the New York City region will lead to such advances without the pursuit of similar objectives in other US cities and other countries and without extensive sharing of genomic and clinical data. Consequently, we offer this preliminary report of our experience to date with P-1000 in the hope that investigators in other regions of the USA and other countries will be encouraged to develop similar programs, to make use of the WGS data that we will ultimately provide, to use such projects as platforms for training students and fellows in the conduct of research on ethnicity, medical genomics, and health disparities, and to confer with us about more equitable ways to pursue cancer science through engagement with populations of patients who more broadly represent the genetic diversity of our species.

Acknowledgments

Funding and other external support for P-1000 were provided by the Mark Foundation for Cancer Research, Illumina, Inc., Cold Spring Harbor Laboratory, Northwell Health, Columbia University, Weill-Cornell Medicine, New York Community Trust, Weslie Janeway, Ben and Donna Rosen, and the Zuckerman Family Fund.

Declaration of interests

No interests are declared.

Resources

[i]www.nygenome.org/research-areas/very-rare-cancer-consortium-2

[ii]www.nytimes.com/interactive/2015/07/08/us/census-race-map.html

[iii]www.coriell.org/1/NHGRI/Collections/1000-Genomes-Collections/1000-Genomes-Project

Supplemental information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.trecan.2021.11.005

[1]New York Genome Center, New York, NY, USA
[2]Weill-Cornell Medicine, New York, NY, USA

*Correspondence:
nrobine@nygenome.org (N. Robine) and
varmus@med.cornell.edu (H. Varmus).

https://doi.org/10.1016/j.trecan.2021.11.005

References

1. Hoadley, K.A. *et al.* (2018) Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.e6
2. AACR Project GENIE Consortium (2017) AACR Project GENIE: powering precision medicine through an international consortium. *Cancer Discov.* 7, 818–831
3. Tate, J.G. *et al.* (2019) COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 47, D941–D947
4. Huang, K.-L. *et al.* (2018) Pathogenic germline variants in 10,389 adult cancers. *Cell* 173, 355–370.e14
5. Yuan, J. *et al.* (2018) Integrated analysis of genetic ancestry and genomic alterations across cancers. *Cancer Cell* 34, 549–560.e9
6. Carrot-Zhang, J. *et al.* (2020) Comprehensive analysis of genetic ancestry and its molecular correlates in cancer. *Cancer Cell* 37, 639–654.e6
7. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) Pan-cancer analysis of whole genomes. *Nature* 578, 82–93
8. Priestley, P. *et al.* (2019) Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* 575, 210–216
9. Arora, K. *et al.* (2019) Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. *Sci. Rep.* 9, 19123
10. Kautto, E.A. *et al.* (2017) Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. *Oncotarget* 8, 7452–7463
11. Chakravarty, D. *et al.* (2017) OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* 2017 PO.17.00011
12. Alexander, D.H. *et al.* (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664
13. Byrska-Bishop, M. *et al.* (2021) High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* Published online November 10, 2021. https://doi.org/10.1101/2021.02.06.430068