Oncology and Informatics – Review

Oncology

Oncology
DOI: 10.1159/000493192

Received: June 25, 2018
Accepted: July 25, 2018
Published online: October 26, 2018

Mountains and Chasms: Surveying the Oncogenomic Publication Landscape

Paula Carrio-Cordo Michael Baudis

Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

Keywords
Cancer genomics · Copy number alteration · Comparative genomic hybridization · Bioinformatics

Abstract
Cancers arise from the accumulation of somatic genome mutations, with varying contributions of intrinsic (i.e., genetic predisposition) and extrinsic (i.e., environmental) factors. For the understanding of malignant clones, precise information about their genomic composition has to be correlated with morphological, clinical, and individual features in the context of the available medical knowledge. Rapid improvements in molecular profiling techniques, the accumulation of a large amount of data in genomic alterations in human malignancies, and the expansion of bioinformatic tools and methodologies have facilitated the understanding of the molecular changes during oncogenesis, and their correlation with clinicopathological phenotypes. Far beyond a limited set of "driver" genes, oncogenomic profiling has identified a large variety of somatic mutations, and whole-genome sequencing studies of healthy individuals have improved the knowledge of heritable genome variation. Nevertheless, the main challenges arise from the skewed representation of individuals from varying population backgrounds in biomedical studies, and also through the limited extent in which some cancer entities are represented in the scientific literature. Content analyses of oncogenomic publications could provide guidance for the planning and support of future studies aiming at filling prominent knowledge gaps.

Introduction

Cancers as Genomic Diseases

Cancers are based on the accumulation of genomic mutations, leading to the transformation of somatic cells into a malignant clone expressing the characteristic "Hallmarks of Cancer" [1]. Different types of cancers show varying types of overall mutation patterns, which may allow identification of molecular subsets beyond traditional diagnostic classifications [2, 3] and can be utilized for prognostic risk assessment and clinical decision making [4, 3].

While the majority of mutations emerge during an individual's lifetime ("somatic" mutations), the risk for developing a malignant disease can be influenced by inherited ("germline") genome variations. Some mutations predisposing to specific malignancies have been identi-

KARGER

E-Mail karger@karger.com
www.karger.com/ocl

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences, University of Zurich
Winterthurerstrasse 190
CH–8057 Zurich (Switzerland)
E-Mail m@baud.is

fied due to high penetrance and apparent familial inheritance pattern [5–7]. However, the interaction of multiple genetic variants with lifetime cancer risk is still poorly understood, reflecting part of the "missing heritability" [8] of complex diseases.

Germline variants may correlate with the population background of individuals and be associated – by approximation – with their geographical origin. Although socioeconomic factors differ in their geographic distribution and contribute to disease incidence and mortality in general, the strong association of several inherited single nucleotide variations with specific cancers motivates a more thorough search for a heritable influence on somatic variation patterns. Differences in the inherited genomic background may be correlated with the amount and types of acquired mutations during cancer development [9, 10], which has implications for understanding the molecular behavior of the tumors as well as on the treatment options for patients [11, 12].

Oncogenomic Screening Techniques

The possibility of alterations of a "heritable agent" in the etiology of cancer had been proposed long before the description of DNA as the molecule of genetic inheritance, but was met with skepticism in its early days, as expressed in this review of Theodor Boveri's work from 1914 [13]):

> ... as well as for its impracticability, it is probable that the hypothesis will not be favorably received by the medical profession.

One of the reasons for early skepticism of chromosomal changes as the basis for cancer development was the impracticability of studying them in humans. However, the development of chromosomal preparation and staining techniques led to an interest in studying the chromosomal composition of neoplastic cells, starting with hematologic malignancies [14, 15] as well as solid tumors [16]. Over the next decades, the field of cancer cytogenetics produced a huge number of studies about chromosomal abnormalities in cancer; currently, the "Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer" reports 68,379 individual cases [17].

While cytogenetic banding can describe "phenotypic" chromosomal observations without analysis of the involved sequence alterations, these observations could be associated with mapped positions of tumor-associated genes [18] or guide their identification [19]. Major progress came from the use of sequence-specific probes using in situ hybridization [20, 21], especially after the introduction of fluorescence in situ hybridization [22, 23] and the delineation of chromosomal fragments in cancer karyotypes using chromosomal "painting" techniques [24, 25]. However, analysis by those technologies required access to karyotypes from dividing cancer cells or was limited to specific measurements, thereby limiting the utility for the discovery of unknown aberrations.

The first whole-genome molecular cytogenetic technology not requiring access to living tumor cells was comparative genomic hybridization (CGH [26, 27]), a reverse in situ hybridization technique in which labeled whole-genomic tumor DNA is hybridized to a matrix of normal human metaphase chromosomes. CGH represented a semi-quantitative analysis of DNA along the whole genome and importantly allowed the use of DNA extracted from source materials such as frozen and archival tissue [28]. While the spatial resolution of chromosomal CGH was especially limited for genomic deletions [29], the analysis of genomic imbalances in neoplasias not amenable to in vitro culture detected unexpected types of genomic alterations [30, 31] and disease-related patterns [32–36].

A major advancement for hybridization-based genomic profiling was the replacement of the hybridization substrate by thousands of defined DNA probes spotted on glass slides. Such "array" or "matrix" CGH experiments (aCGH [33, 36]) permitted the direct assignment of altered sequences. Furthermore, oligonucleotide-based "SNP" arrays, developed for genetic polymorphism profiling [37], were shown to be suitable for copy number profiling [38] and became the predominant genome profiling technology in cancer analysis.

In the last decade, "next generation" sequencing technologies (NGS) have been applied to genome screening experiments in cancer, both for the analysis of whole genomes (whole-genome sequencing, WGS) as well as for whole-exome sequencing (WES) [39]. In addition to detecting single nucleotide variations and other spatially limited sequence variants, the read data from NGS analyses can be used to derive structural variation data, such as regional copy number imbalances [40, 41].

Bioinformatics in Genome Screening

Since the rapidly accumulating biological data is both complex and extensive, bioinformatic procedures are required as enabling technologies for data processing, warehousing, and annotation as well as the biological interpretation of observations and measurements. Over the last decades, specialized areas of bioinformatics have emerged with a focus on, for instance, image analysis, data visualization, systems biology, text mining, and "multiomics,"

Table 1. Characteristics of different genomic screening techniques

1st application report	cCGH, 1992	Genomic arrays, 1997	WES, 2008	WGS, 2008
Genomic resolution	Chromosomal bands = few megabases	Mostly in the 100-kb range	Single bases (2% of the genome)	Single bases
Target identification	Surrogate (position)	"Semidirect" (segmentation spanning probes)	Direct quantitative and qualitative	Direct quantitative and qualitative
Balanced structural	No	No (exceptions)	Depending on position	Yes
Available data	>20,000 cases through Progenetix	Raw (e.g., GEO, arrayExpress) and annotated arrayMap data	Limited (controlled, e.g. TCGA, ICGC)	Limited (controlled, e.g., PCAWG)
Predominant data format	ISCN	Raw; depends on bioinformatics	VCF files	VCF files
Bioinformatics	Image segmentation, densitometry	Value segmentation, background subtraction	Alignment, base quality recalibration, variant calling	Alignment, base quality recalibration, variant calling

with major repercussions for the biomedical community and the field of personalized health.

With a focus on genomic profiling data (Table 1), different bioinformatic approaches are applied depending on the genomic screening technique and target of the analysis. Whereas hybridization-based technologies have important dependencies on image analysis and signal segmentation technologies, a core technique in the processing of NGS data is in the assembly of nucleic acid and protein sequences [42] and mapping of those sequences to reference genomes using a variety of sequence similarity detection algorithms (e.g., Smith-Water man, BLAST, Hidden Markov Models). Further methods, tools, and repositories are continuously being created for the identification and functional assessment of sequence variants.

Although great advances in cancer profiling data analysis have been driven by bioinformatics, a main challenge remains in the integration of data from different sources and technologies. Unfortunately, an extraordinary share of bioinformatic efforts has to be diverted towards data integration, i.e., the mining and harmonization of molecular and metadata, from a vast number of different file formats, data interfaces, and annotation styles.

Published Cancer Genome Screening Studies

In the time since the first application of CGH to screen cancer samples for genomic copy number imbalances, a large number of oncogenomic studies have been published, both in case reports as well as in large studies covering more than 1,000 samples [43, 44]. While the studies considered here were using molecular-cytogenetic and genome sequencing based on different technologies and varying sensitivity and spatial resolution, they all provide a whole-genome read-out for genomic copy number imbalances without selectively targeting specific genome elements.

For our discussions considering the "Oncogenomic Publication Landscape" we will focus on studies of whole-genome, molecular screening techniques using tumor DNA as starting material, including chromosomal (cCGH) and array CGH (aCGH, including single-color oligonucleotide and SNP arrays), as well as WES and WGS. We will use data from existing repositories to highlight biases in published cancer genome screening data, both regarding the representation of disease entities as well as the geographic provenance. For this discussion we will consider the different technologies as "equivalent by intent" – i.e., whole-genome cancer variation profiling – and not with respect to differences in the detection sensitivity or added data qualities beyond structural variation profiling.

Most of the following observations are based on data collected for the Progenetix (progenetix.org [45]) and arrayMap (arraymap.org [46]) resources. Although these curated data repositories cannot provide an exhaustive image of all research in the area, the massive amount of data accumulated there can deliver a representative snapshot of the field, to encourage discussions about study targets and data trajectories.

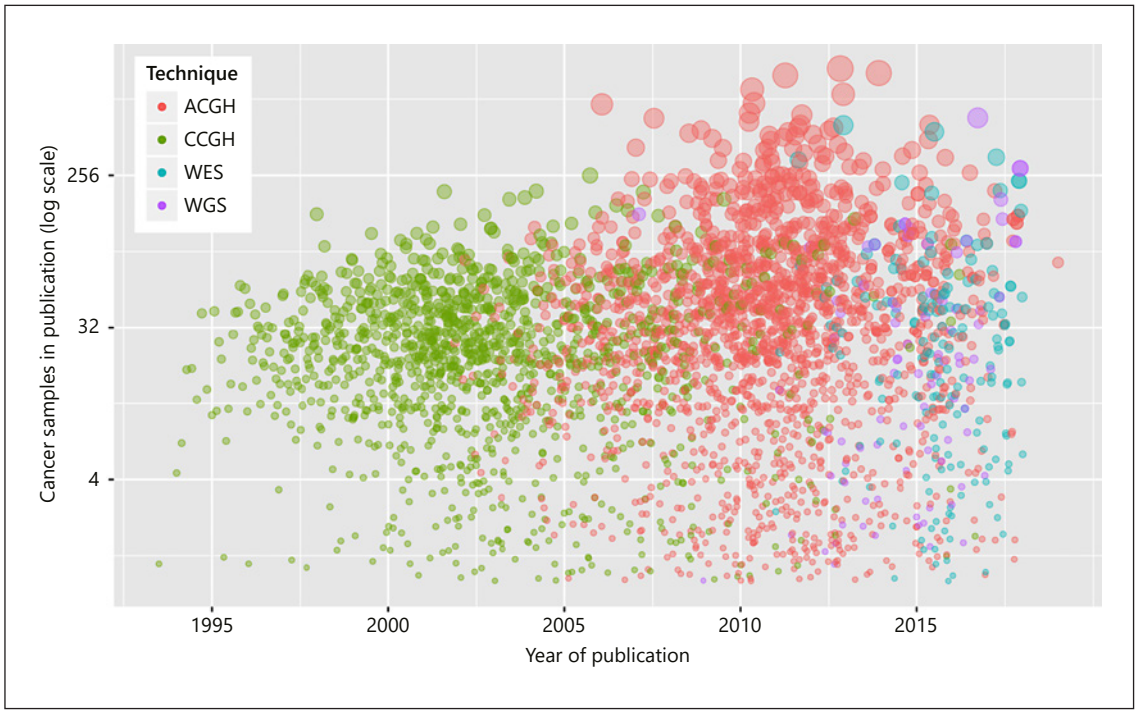

Fig. 1. Publication statistics for cancer genome screening studies. The graphic shows our assessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole-exome and whole-genome sequencing). For the years 1993–2018, we found 3,078 publications reporting 150,000 individual samples in single series from 1 to more than 1,000 samples. The *y* axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

The Progenetix website was established in 2001 [45] to collect and represent data from published CGH studies for comparative meta-analyses of genomic copy number profiles. In identifying data suitable for the resource, over the years a main feature became the general tracking of publications about cancer genome screening studies, independent of the accessibility of the raw data itself. Data attributes for each publication registered in Progenetix and relevant for the discussions are, for example, the number of cancer samples per technological category (cCGH, aCGH, WES, WGS), the geographic provenance of the samples (approximated by the location of the study's corresponding author), as well as the "cancer type" reported. Where available, sample-specific copy number imbalance data is collected and represented in various formats [47].

In contrast to the *Progenetix* resource, the *arrayMap* cancer genome repository represents genome profiling data through mining and re-processing of genomic array data, currently including more than 260 platform types with the minimum requirement of whole-genome probe level representation. As in the case for Progenetix, main features are data curation and annotation in standardized formats, as well as the graphical representation of genome variation data [46].

While the main reason for individual genome screening analyses is the discovery of genome variants without a priori target selection, an added benefit lies in the possibility to assemble large datasets for meta-analyses of cancer-related genome variant frequencies and patterns. Such datasets enable comparative studies of driver gene involvement (e.g., MYCN, BCL2, TP53, HER2, CDKN2A/B, or BRAF) across different cancer types. Also, since many potential gene targets in genomic regions with recurring copy number alterations across cancer types still remain to be identified, events such as the recurrence of focal genome alterations have been argued to represent consequences of strong selection on limited structural rearrangement events during cancer evolution [48, 49] and can be used to pinpoint candidate oncogene involvement based on statistical analyses [50].

The integration of cancer genomic data across studies can help to define the genetic landscape of different can-

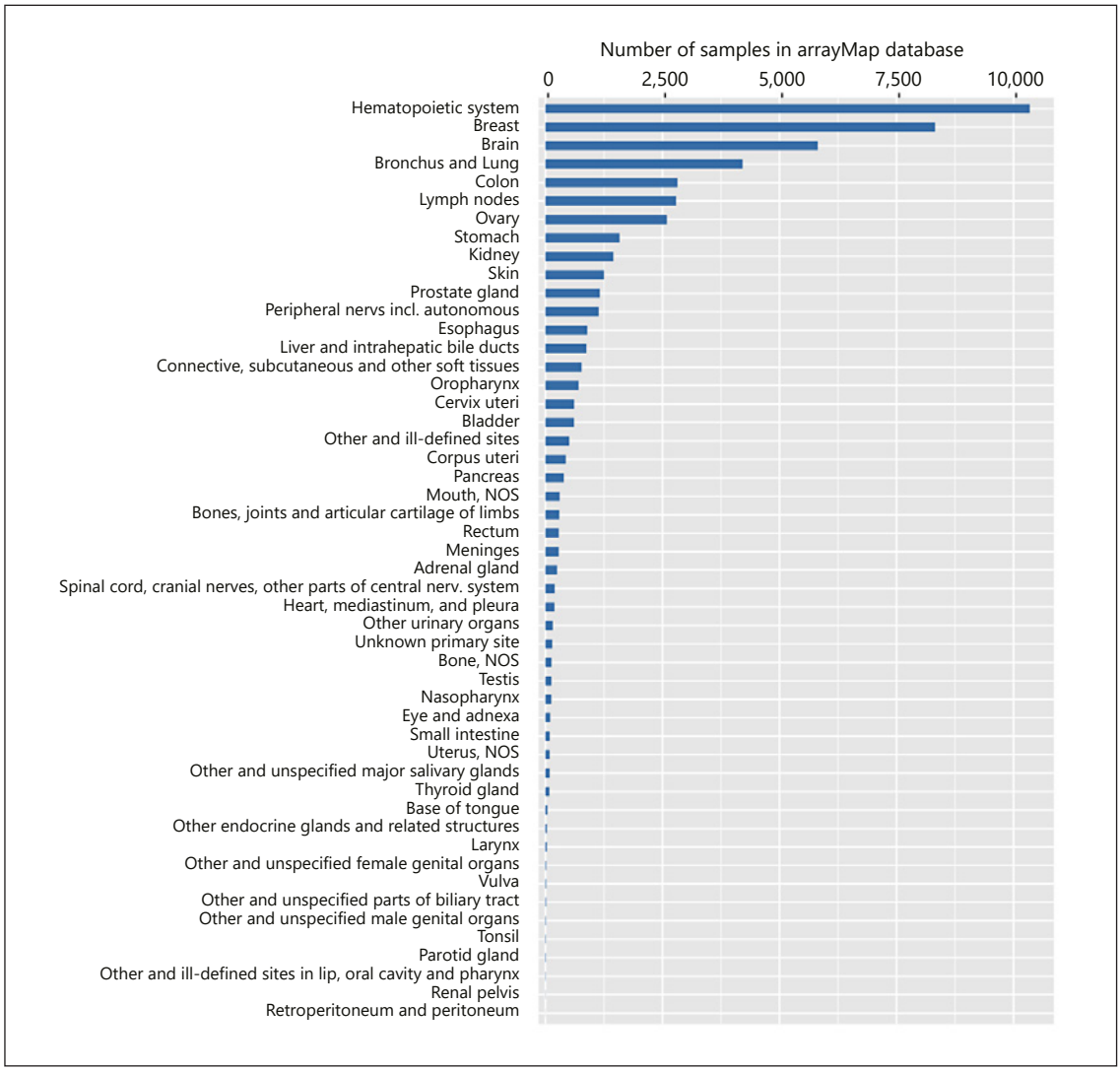

Fig. 2. Distribution of the 50 most studied cancer times based on entities represented in arrayMap by ICD-O-3 topography (i.e., organ site)

cer types. As an example, in a study combining genomic data of breast and colorectal cancers, 189 genes were identified to contribute to neoplastic processes. These genes were previously unknown to be modified in cancers but they reveal certain cancer-specific patterns by integrative analysis [51]. This kind of integrative approach is beneficial not only to find novel gene targets, but also to discover the general patterns across various cancer types.

As a resource for the identification of existing reports for specific cancer types as well as for the assembly of meta-analysis studies, the Progenetix resource currently provides metadata on more than 3,000 articles published between 1993 and 2018, representing 36,496, 102,009, 7,023, and 3,343 individual cancer samples analyzed by cCGH, aCGH, WES, and WGS, respectively. Figure 1 displays the temporal distribution of these publications, with indications for the number of presented samples and used technologies. While these numbers have a certain temporal lag – both due to delay between data production and publication and delays in identification and annotation of the respective articles – one can observe the general trends to move towards newer technologies and higher sample numbers per published study, with NGS-based studies increasingly replacing hybridization-based analyses (so far with lower, but increasing, numbers per study).

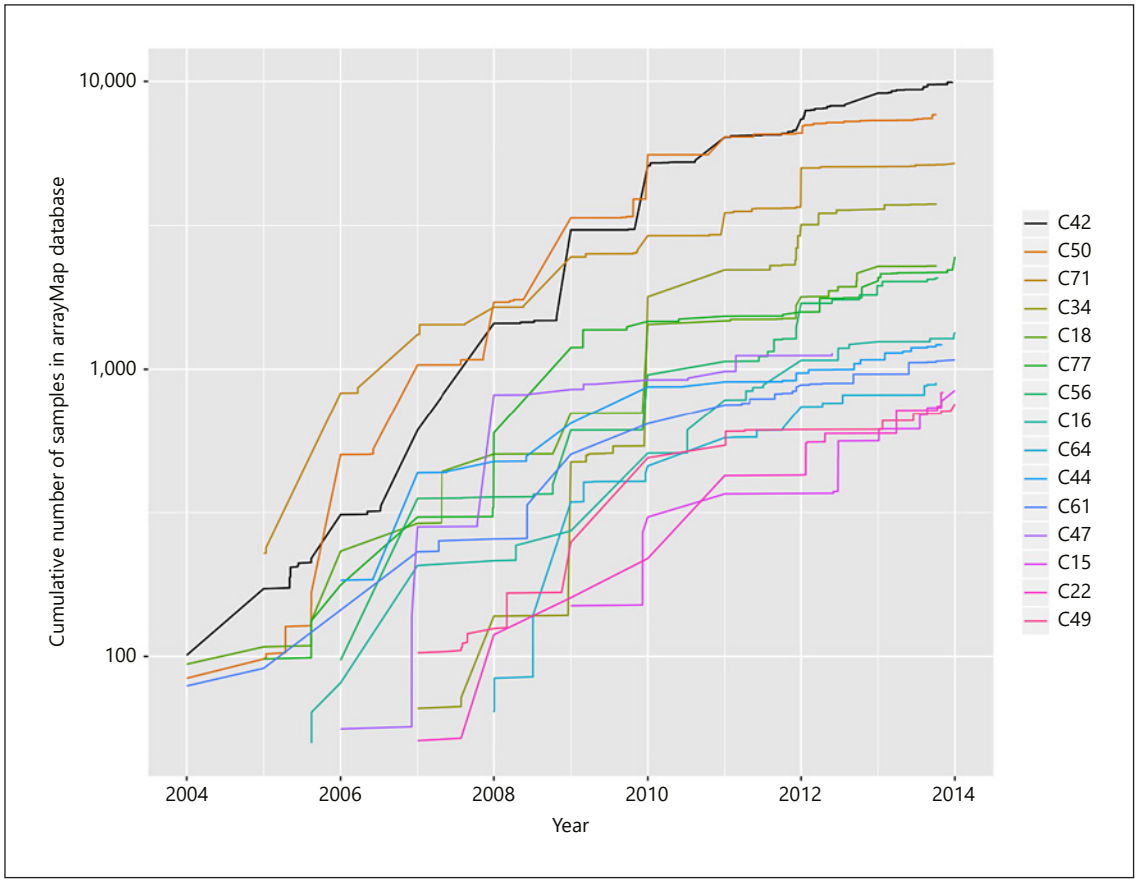

Fig. 3. Cumulative number of samples of the 15 most represented cancer types by ICD-O topography codes in the arrayMap database, over a 10-year period (sample numbers in logarithmic scale).

Representation of Diagnostic Classes

While the overview of the oncogenomic publication space gives some indication of the overall amount of data being produced in research studies, these estimates do not provide information about data produced for different cancer types. For an approximation of the availability of diagnosis-matched genome profiles, one can utilize resources which provide per sample metadata, with annotations mapped to uniform classification systems. Such resources can consist of collaborative projects, such as The Cancer Genome Analysis project (TCGA [52]) or the International Cancer Genome Consortium (ICGC [53]), where many individual research groups contribute molecular profiling and metadata of different tumor types in a coordinated fashion, or in curated data resources.

For our arrayMap resource (arraymap.org [46]) we utilize different primary data sources such as the EBI arrayExpress [54], publication supplements, and user-provided data. However, arrayMap data chiefly reflects the content of the NCBI GEO resource [55] for cancer datasets from suitable genomic array platforms. To date, 267 different platforms and 901 experimental series are available for copy number alteration arrays. As result of the continuous data integration performed through a semi-automated data processing and annotation pipeline [46], at this time a total of 250 morphologies from 94 distinct topographies have been annotated according to ICD-O-3 [56].

As seen in Figure 2, the vast majority of the samples from arrayMap are from hematologic neoplasias, breast cancer, brain tumors, lung and bronchus carcinomas, and colorectal cancers with a representation of 25 (20 + 5 for other NHL), 16, 11, 8, and 5%, respectively. When including the year of publication, we observe that the relative contributions are approximately maintained within the 10-year period (Fig. 3). While the granularity of diagnostic assignments may differ between studies, it is striking that more than half of the data is derived from 4% of 94 registered cancer sites.

Table 2. Numbers of publications and associated samples in the Progenetix article registry, separated for geographic regions

a 1st application report

	Chromosomal CGH, *n* = 199,226	Genomic, arrays, *n* = 199,733	WES, *n* = 200,834	WGS, *n* = 200,835
Resolution	Chromosomal bands = few megabases	Mostly in the 100-kb range	Single bases (2% of the genome)	Single bases
Target identification	Surrogate (position)	"Semidirect" (segmentation spanning probes)	Direct quantitative and qualitative	Direct quantitative and qualitative
Balanced structural	No	No (exceptions)	Depending on position	Yes
Available data	>34,000 cases through Progenetix	Raw (e.g., GEO, arrayExpress) and annotated arrayMap data	Limited (controlled, e.g., TCGA, ICGC)	Limited (controlled, e.g., PCAWG)
Predominant data format	ISCN = static	Raw; depends on bioinformatics	VCF files	VCF files
Bioinformatics	Image segmentation, densitometry	Value segmentation, background subtraction	Alignment, base quality recalibration, variant calling	Alignment, base quality recalibration, variant calling

b Publications

Samples per technique					(Sub-) continent
	cCGH	aCGH	WES	WGS	
3	58	33	0	0	Africa
46	225	1,444	303	288	Australia
465	6,132	10,736	1,534	579	East Asia
28	564	1,324	45	0	South Asia
31	720	209	0	0	Western Asia
1,619	24,070	47,251	1,912	1,465	Europe
833	4,680	39,352	2,888	936	North America
29	208	230	2	1	South America

A selection bias regarding cancer types is also apparent when comparing study representation (arrayMap and TCGA) with the respective incidences. While breast cancer cases represent 15.3% of all cancers [57], in arrayMap 15.8% (9.70% TCGA) of samples were identified as representing a type of breast carcinoma. However, prostate cancer accounting for 9% of all new cases is underrepresented with only 2.21% (4.41%) of study samples. Bladder cancer, which accounts for 4.7% of all new cancer cases, has 1.16% (3.66%) of the sample representation. Thyroid cancer has 3.1% incidence with 0.16% (4.48%) of samples, and larynx carcinoma has 0.8% incidence with 0.05% of samples (1.11% TCGA).

Moreover, whereas some of the most studied cancers have low mortality rates such as breast cancer with almost 90% survival after 5 years, special mention should be made of those entities underrepresented and with high mortality. For instance, pancreas cancers have 0.75% of samples in arrayMap (1.63% TCGA) but 3.2% of all new cases have an 8.5% 5-year survival rate. While esophagus cancer is proportionally well represented (1.70/1.63% of samples for 1% of all new cancers), it remains poorly understood with a 5-year survival rate still at 19.2%.

Overall, cancer genome publications reflect the preferred analysis of frequent cancers with some apparent biases, while being limited in the representation of rare tumor types. Multiple factors could explain biases in cancer type selections: lack of general interest and major problematic assembly of biosamples for rare cancer types, allocation of research funding for specific cancer types

(e.g., breast cancer) due to public perception and advocacy, lack of availability of tissue samples due to technical difficulties in sample extraction and processing, or ethical and legal implications regarding patient privacy in sample sharing for genomic analysis [58–60]. To relieve these disparities, global and efficient actions should be taken. While the current tendency is indeed to study cancer types with high incidences, the study of rare entities could dramatically increase our knowledge of cancer biology.

Geographies of Published Studies

A number of studies remark on disparities in cancer incidence, prevalence, and mortality related to ethnicity and geographic origin [61, 62]. Two general classes of factors have been found to contribute to these disparities: (a) environmental factors through different types and levels of exposure related to local or regional geographical origin, and (b) population-specific variation in genomic variants with influence of heritable contributions on cancer development.

Although, many studies relate the influence of geographic pattern incidences with environmental factors such as pollution levels, intensity of UV radiation, or exposure to infectious agents [61], the contribution of population-specific biases in cancer promoting genome variants is less well defined. Some relevant studies in the area have shown the BRCA1 gene as a population-specific bias in some homogeneous groups compared to outbred reference populations [63]. In the assembly of a meta-resource for oncogenomic publications, the contact information of the corresponding authors represents an important piece of information, e.g., for facilitating the contacting of study authors by the resource's users, for follow-up questions, or access to detailed study information or source data. However, this information can also be used as proxy to provide quantitative representation of the study content with relation to geographic provenance, leading to some interesting observations.

The geographic origin mapping of more than 3,000 publications represented in the Progenetix article registry showed large biases regarding the provenance of the published data (Table 2). While the overall preponderance of studies from Europe (1,619) and North America (833) could be expected, the near complete lack of cancer genome screening studies from the African continent was unexpected.

Since cancer development can be influenced by population-related inherited genome variants as well as extrinsic factors related to local environmental exposure and sociocultural practices, it is of paramount importance to include geolocation metadata in the assessment of molecular profiling. However, the real impact of factors correlated with geographic provenance can only be assessed with the availability of sufficient, representative data for a large range of geographies, ethnicities, and environments.

Focusing on the geographic location of the studies, the tendency is, as expected, for developed countries to provide the majority of oncogenomic data (Fig. 4). Most of the published studies are reported from Europe, the USA, China, Japan, Australia, and the Korean Republic. In contrast, only very few studies have been reported from Central and South Asia as well as South America. However, most striking is the near complete absence of cancer genome studies from the African continent in the accessible literature. One can assume that these geographic biases reflect major difficulties in the establishment of technology-driven research in underdeveloped countries – from lack of training of scientists to infrastructure problems for biosample extraction, expense and availability of reagents and technical equipment, as well as computing infrastructure for bioinformatic processing [64].

A more uniform map of the number of genomic studies across the world is of further importance for clinical trials, where attention on the genetic variation across ethnic groups could improve novel therapies and reduce cancer disparities.

GA4GH to the Rescue

While the need for more and diverse studies of cancer genome mutation profiles and their relation to the underlying personal genomes is increasingly being realized, a major obstacle in utilizing the emerging data lies in the high degree of fragmentation and "siloing" of the generated data. Genomics and associated metadata are frequently created as part of research studies with study-specific consent [65] and access restricted to original study collaborators. If mechanisms for outside researchers are in place, they usually require submission of specific project proposals and review through a data access committee and agreement to specific usage conditions. If data then can be accessed, it is in a variety of genome and metadata formats which usually have to be normalized to common encodings.

The core mission of the Global Alliance for Genomics and Health initiative (GA4GH [66]) is to "... enable genomic data sharing for the benefit of human health". Its members address this goal through the improvement of global standards and the creation of tools to securely share genomic data across geographic or institutional

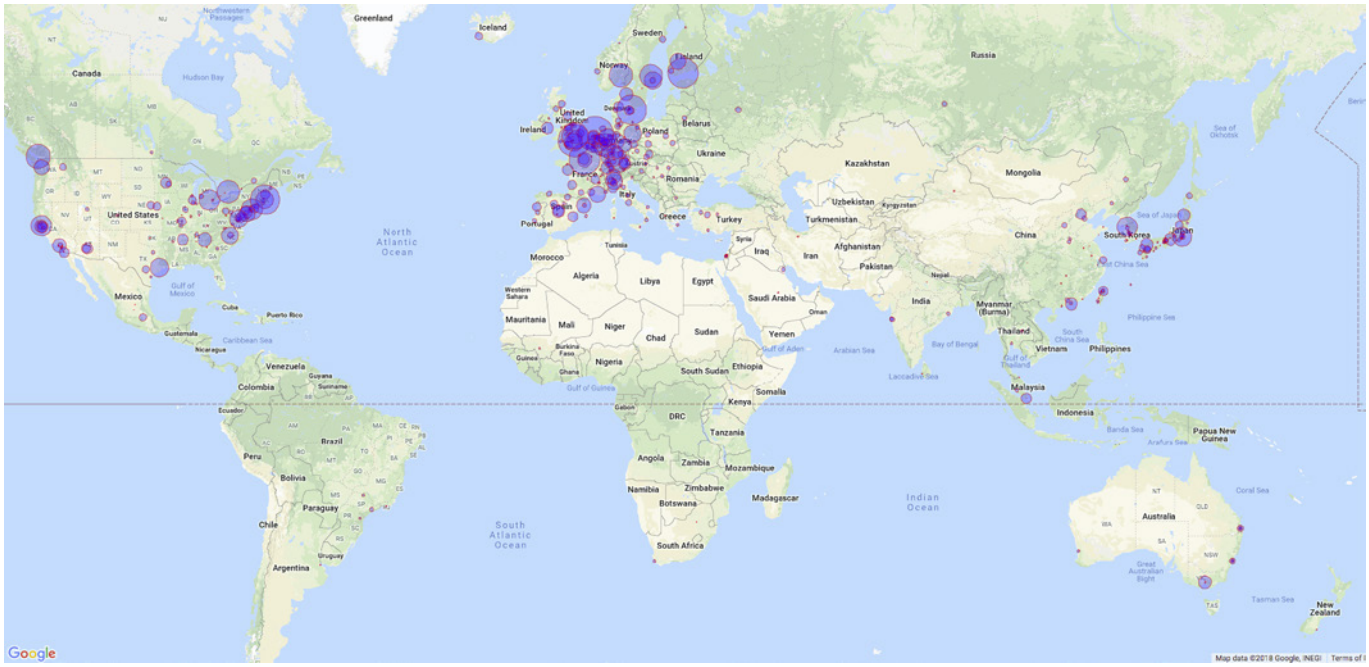

Fig. 4. The map displays the geographic distribution (by corresponding author) of the 102,225 genomic arrays, 36,747 chromosomal CGH, and 7,212 whole-genome/exome-based cancer genome datasets. The numbers are derived from the 3,103 publications registered in the Progenetix database.

boundaries. Formally established in 2014 and with an increasing participation of currently 499 organizational members from 44 countries, the GA4GH has started to shape the public discourse about the benefits of genome-driven research for human health applications, and started to provide guidelines [67], standards, and tool kits to enable secure and ethically responsible data sharing. These activities are based on the work of different "work streams", which interact with existing "driver projects" in the iterative development, testing, and implementation of protocols, standards, and tool kits. The driver projects themselves – such as the "Beacon" [68] project or the "BRCA exchange" [69] – address particular scientific, technical, regulatory, or security-related aspects of federated access to human genomes and related metadata. However, while the development of protocols, tools, and guidelines for the effective sharing of genome-related data is a prerequisite for widening the scope and statistical significance of studies in biomedical genomics, by themselves these efforts alone cannot solve the skewed generation of genome screening data with respect to disease representation and REA (race, ethnicity, and ancestry) provenance. Additionally, having suitable protocols and tools at hand does not guarantee their implementation and use by the providers of the many institutional or national resource providers. These problems can only be addressed in an iterative process, involving coordinative work by organizations such as GA4GH in interaction with national and international policy makers and funders of scientific projects and research infrastructures.

Conclusions

Continuous efforts into the understanding of tumor biology have led to an increasing number of coordinated international projects generating oncogenomic data. This progress has been made possible by the development of genome screening techniques, supported through the rapid advancement in computational hardware and bioinformatic tools. Nowadays, the tight integration of bioinformatics can be considered essential not only for meta-analyses and statistical studies, but also as a necessary element in the execution of all types of molecular analyses and data management pipelines.

However, the ability to use text mining and other bioinformatic tools to create large surveys of existing genome studies now allows us to observe biases in the data

being reported, both with respect to the representation of tumor entities as well as in the general lack of data from large fractions of the world's populations. Impacts of these biases can be suspected in the missing opportunities for insights into particular oncogenetic mechanisms in rare cancer types, and the failure to fulfil the promise of "Precision Medicine" to those patients.

The other type of bias discussed here is the highly limited representation of many human populations – particularly from Africa – in publications reporting data from cancer genome screening analyses. The resulting lack of ethnic diversity will still be a barrier in trying to elucidate molecular events related to specific population backgrounds, thereby possibly missing out on specific therapeutic targets. These biases are not only limited to cancer, with recent data showing that more that 50% of all reported genome variants in the Genome Aggregation Database (gnomAD) are based on European ancestry [70].

Besides the well-known impact of major socioeconomic factors, efforts towards understanding disparities in global cancer incidences and prognostic trajectories should also be directed with the characterization of differences in genetic variation patterns – both inherited polymorphisms and somatic variants in cancer genomes – for large numbers of patients from a variety of population backgrounds. Moreover, researchers should increasingly direct their attention towards rare cancer entities from which the knowledge would dramatically increase in benefit of personalized medicine. Here, one can argue that the limited number of cancer types studied and the low diversity of targeted populations should be addressed through the allocation of financial resources and support of international collaborative efforts.

One important aspect of a truly "global" understanding of every aspect of the impact of inherited and somatic variations on cancer biologies, clinical prognostications, and targeted interventions will be to facilitate data access beyond the current localized data silos and individual publications with, at best, highly fragmented but frequently nonexisting access to genomic and associated metadata. Here, the Global Alliance for Genomics and Health provides a leading effort towards better access to health-related data, beyond individual studies and localized repositories, towards a global network of interacting standards and resources.

Acknowledgments

M.B. wishes to thank the Kavli Institute for Theoretical Physics (KITP), as a guest of which parts of this manuscript were written. Therefore, this research was supported in part by the National Science Foundation (grant No. NSF PHY-1748958). Additional thanks go to the current and former members of the Baudis group at the University of Zurich for their continuing efforts in making cancer genome data accessible.

Disclosure Statement

The authors declare that they have no conflicts of interest.

References

1 Hanahan D, Weinberg RA: Hallmarks of cancer: the next generation. Cell 2011;144:646–674.
2 Lapointe J, Li C, Giacomini CP, Salari K, Huang S, Wang P, Ferrari M, Hernandez-Boussard T, Brooks JD, Pollack JR: Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. Cancer Res 2007; 67:8504–8510.
3 Jones DT, Jager N, Kool M, Zichner T, Hutter B, Sultan M, et al: Dissecting the genomic complexity underlying medulloblastoma. Nature 2012;488:100–105.
4 Vandesompele J, Baudis M, De Preter K, Van Roy N, Ambros P, Bown N, Brinkschmidt C, Christiansen H, Combaret V, Lastowska M, Nicholson J, OMeara A, Plantaz D, Stallings R, Brichard B, Van den Broecke C, De Bie S, De Paepe A, Laureys G, Speleman F: Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma. J Clin Oncol 2005;23:2280–2299.
5 Malkin D, Li FP, Strong LC, Fraumeni JF, Nelson CE, Kim DH, Kassel J, Gryka MA, Bischoff FZ, Tainsky MA: Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science 1990;250:1233–1238.
6 Spirio L, Otterud B, Stauffer D, Lynch H, Lynch P, Watson P, Lanspa S, Smyrk T, Cavalieri J, Howard L: Linkage of a variant or attenuated form of adenomatous polyposis coli to the adenomatous polyposis coli (APC) locus. Am J Hum Genet 1992;51:92–100.
7 Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W: A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266:66–71.
8 Maher B: Personal genomes: the case of the missing heritability. Nature 2008;456:18–21.
9 Deng J, Chen H, Zhou D, Zhang J, Chen Y, Liu Q, et al: Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. Nat Commun 2017;8:1533.
10 Heath EI, Lynce F, Xiu J, Ellerbrock A, Reddy SK, Obeid E, Liu SV, Bollig-Fischer A, Separovic D, Vanderwalde A: Racial disparities in the molecular landscape of cancer. Anticancer Res 2018;38:22352240.
11 Kim J, Sun CL, Mailey B, Prendergast C, Artinyan A, Bhatia S, Pigazzi A, Ellenhorn JD: Race and ethnicity correlate with survival in patients with gastric adenocarcinoma. Ann Oncol 2010;21:152–160.

12 Keenan T, Moy B, Mroz EA, Ross K, Niemierko A, Rocco JW, Isakoff S, Ellisen LW, Bardia A: 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 2015;33:3621–3627. PMID: 26371147.
13 Calkins GN: Zur Frage der Entstehung maligner Tumoren. Science 1914;40:857–859.
14 Nowell PC, Hungerford DA: A minute chromosome in chronic granulocytic leukemia. Science 1960;132:1497.
15 Rowley JD: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 1973;243:290–293.
16 Zang KD, Singer H: Chromosomal constitution of meningiomas. Nature 1967;16:84–85.
17 Johansson B, Mitelman F, Mertens F (eds): Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. 2018. http://cgap.nci.nih.gov/Chromosomes/Mitelman.
18 Le Beau MM, Westbrook CA, Diaz MO, Rowley JD: c-src is consistently conserved in the chromosomal deletion (20q) observed in myeloid disorders. Proc Natl Acad Sci USA 1985; 82:6692–6696.
19 Alitalo K, Schwab M, Lin CC, Varmus HE, Bishop JM: Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. Proc Natl Acad Sci USA 1983;6:1707–1711.
20 Malcolm S, Barton P, Murphy C, Ferguson-Smith MA, Bentley DL, Rabbitts TH: Localization of human immunoglobulin kappa light chain variable region genes to the short arm of chromosome 2 by in situ hybridization. Proc Natl Acad Sci USA 1982;79:4957–4961.
21 Schwab M, Ellison J, Busch M, Rosenau W, Varmus HE, Bishop JM: Enhanced expression of the human gene N-myc consequent to amplification of DNA may contribute to malignant progression of neuroblastoma. Proc Natl Acad Sci USA 1984;15:4940–4944.
22 Lichter P, Cremer T, Borden J, Manuelidis L, Ward DC: Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant DNA libraries. Hum Genet 1988;80:224–234.
23 DC Tkachuk, CA Westbrook, M Andreeff, TA Donlon, ML Cleary, K Suryanarayan, M Homge, A Redner, J Gray, Pinkel D: Detection of bcr-abl fusion in chronic myelogeneous leukemia by in situ hybridization. Science 1990;250:559–562.
24 Speicher MR, Gwyn Ballard S, Ward DC: Karyotyping human chromosomes by combinatorial multi-fluor fish. Nat Genet 1996;12: 368–375.
25 Veldman T, Vignon C, Schrock E, Rowley JD, Ried T: Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping. Nat Genet 1997;15:406–410.
26 Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D: Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992;5083:818–821.
27 Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P: Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. Hum Genet 1993;90:584–589.
28 Speicher MR, du Manoir S, Schrock E, Holtgreve-Grez H, Schoell B, Lengauer C, Cremer T, Ried T: Molecular cytogenetic analysis of formalin-fixed, paraffin-embedded solid tumors by comparative genomic hybridization after universal DNA-amplification. Hum Mol Genet 1993;11:1907–1914.
29 Bentz M, Plesch A, Stilgenbauer S, Dohner H, Lichter P: Minimal sizes of deletions detected by comparative genomic hybridization. Genes Chromosomes Cancer 1998;2:172–175.
30 Werner CA, Dohner H, Joos S, Trumper LH, Baudis M, Barth TF, Ott G, Moller P, Lichter P, Bentz M: High-level DNA amplifications are common genetic aberrations in B-cell neoplasms. Am J Pathol 1997;151:335–342.
31 Knuutila S, Bjorkqvist AM, Autio K, Tarkkanen M, Wolf M, Monni O, Szymanska J, Larramendy ML, Tapper J, Pere H, El-Rifai W, Hemmer S, Wasenius VM, Vidgren V, Zhu Y: DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. Am J Pathol 1998;152:1107–1123.
32 Weber RG, Bostrom J, Wolter M, Baudis M, Collins VP, Reifenberger G, Lichter P: Analysis of genomic alterations in benign, atypical, and anaplastic meningiomas: toward a genetic model of meningioma progression. Proc Natl Acad Sci USA 1997;26:14719–14724.
33 Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P: Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer 1997;4:399–407.
34 Campbell PJ, Stephens PJ, Pleasance ED, OMeara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA: Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40:722729.
35 Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al: DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 2008;456:66–72.
36 Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 1998;2:207–211.
37 Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES: Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.
38 Zhao X, Li C, Paez JG, Chin K, Jnne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M: An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 2004; 64:3060–3071.
39 Meyerson M, Gabriel S, Getz G: Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010;11:685–696.
40 de Ligt J, Boone PM, Pfundt R, Vissers LE, Richmond T, Geoghegan J, O'Moore K, de Leeuw N, Shaw C, Brunner HG, Lup- ski JR, Veltman JA, Hehir-Kwa JY: Detection of clinically relevant copy number variants with whole-exome sequencing. Hum Mutat 2013; 34:1439–1448.
41 Wang W, Sun W, Crowley JJ, Szatkiewicz JP: Allele-specific copy-number discovery from whole-genome and whole-exome sequencing. Nucleic Acids Res 2015;43:e90.
42 Iliopoulos CS, Pissis SP: Algorithms for next-generation sequencing data; in Elloumi M, Zomaya AY (eds): Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data. Wiley, 2013, pp 251–280.
43 Holland DG, Burleigh A, Git A, Goldgraben MA, Perez-Mancera PA, Chin SF, Hurtado A, Bruna A, Ali HR, Greenwood W, Dunning MJ, Samarajiwa S, Menon S, Rueda OM, Lynch AG, McKinney S, Ellis IO, Eaves CJ, Carroll JS, Curtis C, Aparicio S, Caldas C: ZNF703 is a common luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. EMBO Mol Med 2011;3:167180.
44 Northcott PA, Shih DJ, Peacock J, Garzia L, Morrissy AS, Zichner T, et al: Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature 2012;488: 49–56.
45 Baudis M, Cleary ML: Progenetix.net: an online repository for molecular cytogenetic aberration data. Bioinformatics 2001;17:1228–1229.

46 Cai H, Kumar N, Baudis M: arrayMap: a reference resource for genomic copy number imbalances in human malignancies. PLoS One 2012;7:e36944.
47 Cai H, Kumar N, Ai N, Gupta S, Rath P, Baudis M: Progenetix: 12 years of oncogenomic data curation. Nucleic Acids Res 2014; 42:D1055–D1062.
48 Werner CA, Dohner H, Joos S, Trumper LH, Baudis M, Barth TF, Ott G, Moller P, Lichter P, Bentz M: High-level DNA amplifications are common genetic aberrations in B-cell neoplasms. Am J Pathol 1997;151:335–342.
49 Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al: The landscape of somatic copy-number alteration across human cancers. Nature 2010;463:899–905.
50 Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, et al: Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci USA 2007;104:20007–20012.
51 Sjblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE: The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268–274.
52 Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–1068.
53 Hudson TJ; International Cancer Genome Consortium: International network of cancer genome projects. Nature 2010;464:993–998.
54 EMBL-EBI: Embl-EBI ArrayExpress. https://www.ebi.ac.uk/arrayexpress/arrays/A-AFFY-54/ (accessed June 30, 2017).
55 NCBI: NCBI gene expression omnibus. www.ncbi.nlm.nih.gov/geo/ (accessed June 30, 2017).
56 Fritz A, Percy C, Jack A, Sobin LH, Parkin MD (eds): International Classification of Diseases for Oncology (ICD-O), ed 3. World Health Organization, Geneva, 2000.
57 National Cancer Institute: Surveillance, Epidemiology, and End Results Program. 2018. https://seer.cancer.gov/statfacts/.
58 Casali PG: Rare cancers: work in progress in Europe. Ann Oncol 2014;25:914.
59 Bevilacqua G, Bosman F, Dassesse T, Hfler H, Janin A, Langer R, Larsimont D, Morente MM, Riegman P, Schirmacher P, Stanta G, Zatloukal K, Caboux E, Hainaut P: The role of the pathologist in tissue banking: European consensus expert group report. Virchows Arch 2010;456:449–454.
60 Dove ES: Biobanks, data sharing, and the drive for a global privacy governance framework. J Law Med Ethics 2015;43:675–689.
61 Danaei G, Vander Hoorn S, Lopez AD, Murray CJ, Ezzati M; Comparative Risk Assessment collaborating group (Cancers): Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. Lancet 2005;366:1784–1793.
62 Siegel RL, Miller KD, Jemal A: Cancer statistics, 2017. CA Cancer J Clin 2017;67:7–30.
63 Szabo CI, King MC: Population genetics of BRCA1 and BRCA2. Am J Hum Genet 1997; 60:1013–1020.
64 Helmy M, Awad M, Mosa KA: Limited resources of genome sequencing in developing countries: challenges and solutions. Appl Transl Genom 2016;9:15–19.
65 Dyke SO, Philippakis AA, Rambla De Argila J, Paltoo DN, Luetkemeier ES, Knoppers BM, Brookes AJ, Spalding JD, Thompson M, Roos M, Boycott KM, Brudno M, Hurles M, Rehm HL, Matern A, Fiume M, Sherry ST: Consent codes: upholding standard data use conditions. PLoS Genet 2016;12:e1005772.
66 Lawler M, Siu LL, Rehm HL, Chanock SJ, Alterovitz G, Burn J, Calvo F, Lacombe D, Teh BT, North KN, Sawyers CL; Clinical Working Group of the Global Alliance for Genomics, and Health (GA4GH): All the worlds a stage: Facilitating discovery science and improved cancer care through the global alliance for genomics and health. Cancer Discov 2015;5: 1133–1136.
67 Global Alliance for Genomics and Health: Framework for Responsible Sharing of Genomic and Health-Related Data. https://www.ga4gh.org/ga4ghtoolkit/regulatoryandethics/framework-for-responsible-sharing-genomic-and-health-related-data/ (accessed May 7, 2018).
68 ELIXIR: Elixir Beacon. https://www.elixir-europe.org/about-us/implementation-studies/beacons (accessed May 7, 2018).
69 BRCA Exchange: BRCA exchange of the global alliance for genomics and health. http://brcaexchange.org (accessed May 7, 2018).
70 Popejoy AB, Ritter DI, Crooks K, Currey E, Fullerton SM, Hindorff LA, Koenig B, Ramos EM, Sorokin EP, Wand H, Wright MW, Zou J, Gignoux CR, Bonham VL, Plon SE, Bustamante CD: The Clinical Imperative for Inclusivity: Race, Ethnicity, and Ancestry (REA) in Genomics. bioRxiv, 2018. https://www.biorxiv.org/content/early/2018/05/09/317800.