

# Mountains and Chasms - Surveying the Oncogenomic Publication Landscape

Paula Carrio Cordo<sup>1,2</sup> and Michael Baudis<sup>1,2</sup>✉

<sup>1</sup>Institute of Molecular Life Science, University of Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland

<sup>2</sup>SIB, Swiss Institute of Bioinformatics, Winterthurerstrasse 190, 8057, Zurich, Switzerland

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 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 clinico-pathological 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, main challenges arise from the skewed representation of individuals from varying population backgrounds in biomedical studies, and also through the limited extend 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.

Cancer genomics, CNA, CGH, bioinformatics

Correspondence: mbaudis@imls.uzh.ch

## 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 to identify molecular subsets beyond traditional diagnostic classifications (2, 3) and can be utilised for prognostic risk assessment and clinical decision making (3, 4).

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 identified due to high penetrance and apparent familial inheritance pattern (5–7). However, the interaction of multiple genetic variants on lifetime cancer risk are still poorly understood, reflecting part of the "missing heritability" (8) of complex diseases.

Germline variants may correlate to the population background of individuals and be associated - by approximation

- with their geographical origin. Although socio-economic factors differ in their geographic distribution and contribute to disease incidence and mortality in general, the strong association of several inherited single nucleotide variations (SNV) 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 to the amount and types of acquired mutations during cancer development (9, 10), which has implications for understanding the molecular behaviour of the tumours as well as on the treatment options for patients (11–13).

**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 scepticism in its early days, as expressed in this review of Theodor Boveri's work from 1914(14):

... 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 scepticism of chromosomal changes as basis for cancer development was the impracticability to study them in humans. However, the development of chromosomal preparation and staining techniques lead to an interest in studying the chromosomal composition of neoplastic cells, starting with hematologic malignancies(15, 16) as well as solid tumors(17, 18). 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(19). 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 (24) or guide their identification (25). Major progress came from the use of sequence-specific probes using in-situ hybridisation (26, 27), especially after the introduction of Fluorescence In-Situ Hybridisation (FISH (28, 29)) and the delineation of chromosomal fragments in cancer karyotypes using chromosomal "painting" techniques (30, 31). However, analysis by those technologies required access to karyotypes from dividing cancer cells or was limited to specific measurements, thereby limiting the utility for discovery of unknown aberrations.

**Table 1.** Characteristics of different genomic screening techniques.

	chromosomal CGH	genomic arrays	WES	WGS
<b>1st application report</b>	1992(20)	1997(21)	2008(22)	2008(23)
<b>resolution</b>	chromosomal bands = few megabases	mostly in the 100kb range	single bases ( 2% of the genome)	single bases
<b>target identification</b>	surrogate (position)	"semidirect" (segmentation spanning probes)	direct quantitative and qualitative	direct quantitative and qualitative
<b>balanced structural</b>	no	no (exceptions)	depending on position	yes
<b>available data</b>	>20,000 cases through <i>Progenetix</i>	raw (GEO, arrayExpress) & annotated <i>arrayMap</i> data	limited (controlled, e.g. e.g. TCGA, ICGC)	limited (controlled, e.g. PCAWG)
<b>predominant data format</b>	ISCN = static	raw; depends on bioinformatics	VCF files	VCF files
<b>bioinformatics</b>	image segmentation, densitometry	value segmentation, background subtraction	alignment, base quality recalibration, variant calling	alignment, base quality recalibration, variant calling

The first whole-genome molecular cytogenetic technology not requiring access to living tumor cells was Comparative Genomic Hybridisation (CGH(20, 30, 32)), a reverse in-situ hybridisation technique in which labeled whole-genomic tumor DNA is hybridised to a matrix of normal human metaphase chromosomes. CGH represented the semi-quantitative analysis of DNA along the whole genome and importantly allowed the use of DNA extracted from a variety of source materials, including frozen and archival tissue(33, 34). While the spatial resolution of chromosomal CGH was limited especially regarding genomic deletions(35), the analysis of genomic imbalances in neoplasias not amenable to in vitro culture detected unexpected types of genomic alterations(36–38) and disease-related patterns(39).

A major advancement for hybridisation based genomic profiling was the replacement of the hybridisation substrate by thousands of defined DNA probes spotted on glass slides. Such "array" or "matrix" CGH experiments (aCGH(21, 40, 41)) permitted the direct assignment of altered sequences. Furthermore, oligonucleotide based "SNP"-arrays, developed for genetic polymorphism profiling(42), were shown to be suitable for copy number profiling(43) 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 (WGS) as well as for whole exome sequencing (23, 44, 45). In addition to detecting single nucleotide variations (SNV) 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 (46, 47).

**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, specialised areas of bioinformatics have emerged with focus on, for instance, image analysis, data visualization, systems biology, text mining and "multi-omics", with major repercussions for the biomedical community and the field of personalised health.

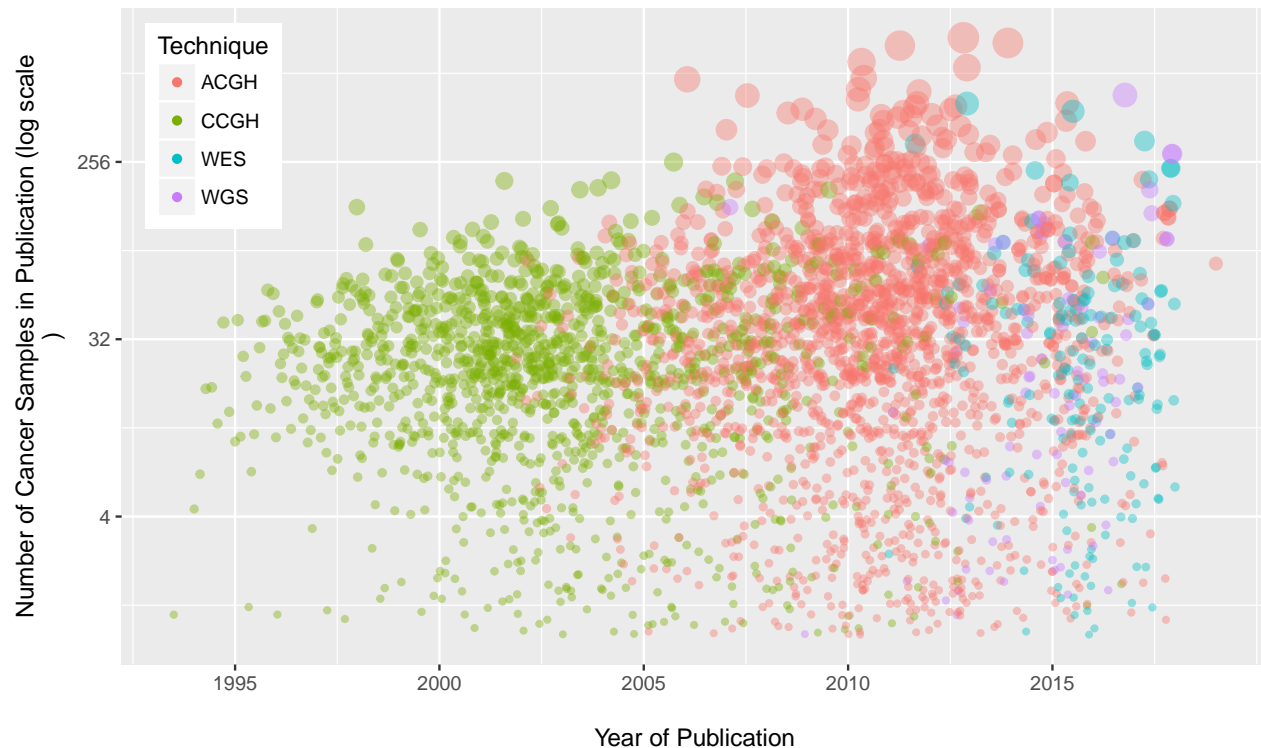
With a focus on genomic profiling data (table 1), different bioinformatics approaches are applied depending on genomic screening technique and target of the analysis. Whereas hybridisation based technologies have important dependencies on image analysis and signal segmentation technologies, a core technique in the processing of NGS data is in the assembly of nucleic acid and protein sequences (48) and mapping of those sequences to reference genomes using a variety of sequence similarity detection algorithms (e.g. Smith-Waterman, 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 harmonisation of molecular and metadata, from a vast number of different file formats, data interfaces and annotation styles.

Published Cancer Genome Screening Studies

## 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 amount of oncogenomic studies has been published, both in case reports as well as in large studies covering more than 1000 samples (49, 50). 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 tar-



**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 genome sequencing). For the years 1993-2018, we found 3078 publications reporting 150000 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

getting 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 whole exome and genome sequencing (WES, 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 with view of 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](http://progenetix.org) (51)) and arrayMap ([arraymap.org](http://arraymap.org) (52)) 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. The Progenetix website was established in 2001 (51), 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 e.g. 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 (53).

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 standardised formats, as well as the graphical representation of genome variation data (52).

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 CNA across cancer types still remain to be identified, events such as the recurring occurrence of focal genome alterations have been argued to repre-

sent consequences of strong selection on limited structural rearrangement events during cancer evolution (54, 55) and can be used to pinpoint candidate oncogene involvement based on statistical analyses (56).

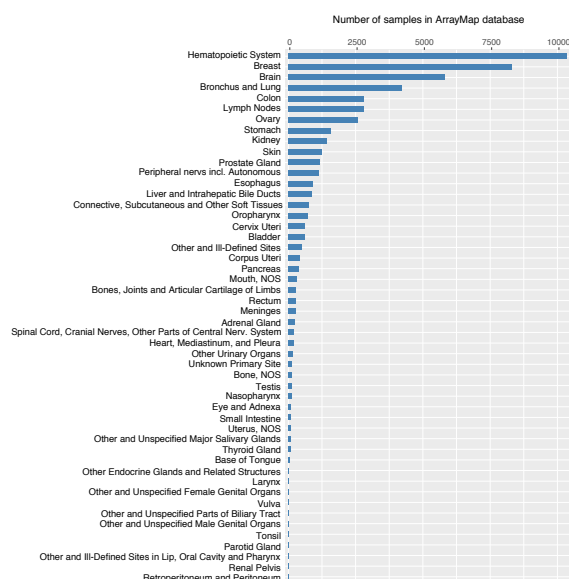
The integration of cancer genomic data across studies can help to define the genetic landscape of different cancer types. As example, in a study combining genomic data of breast and colorectal cancers, 189 genes were identified to contribute to neoplastic processes. These genes were previously unknown to be modified in cancers but they reveal certain cancer-specific patterns by the integrative analysis (57). 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 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 about more than 3000 articles published between 1993 and 2018, representing 36496, 102009, 7023 and 3343 individual cancer samples analysed 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 hybridisation based analyses (so far with lower, but increasing, numbers per study).

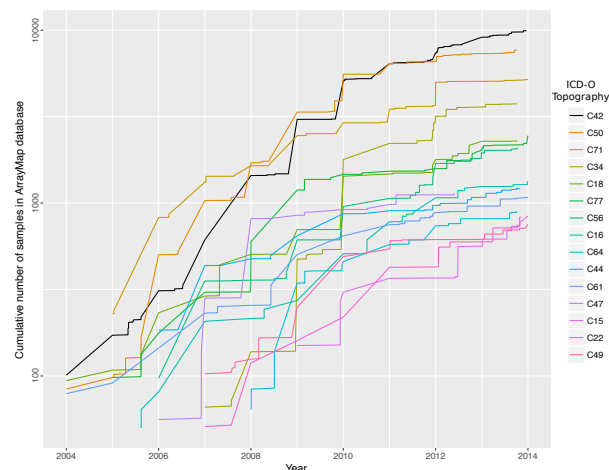
**Representation of diagnostic classes.** While the overview of the oncogenomic publication space gives some indication about the overall amount of data being produced in research studies, these estimates do not provide information about data produced for different cancer types. For an approximation of the availability of diagnosis-matched genome profiles, one can utilise 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 (58)) or the International Cancer Genome Consortium (ICGC (59)), 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*(52)) we utilise different primary data sources such as EBI's *arrayExpress*(60), publication supplements and user provided data. However, *arrayMap* data chiefly reflects the content of NCBI's GEO resource(61), for cancer data sets from suitable genomic array platforms. To date, 267 different platforms and 901 experimental series are available for CNA arrays. As result of the continuous data integration performed through a semi-automated data processing and annotation pipeline(52), at this time a total of 250 morphologies from 94 distinct topographies have been annotated according to ICD-O 3(62)).

As seen in figure 2 the vast majority of the samples from ar-



**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)

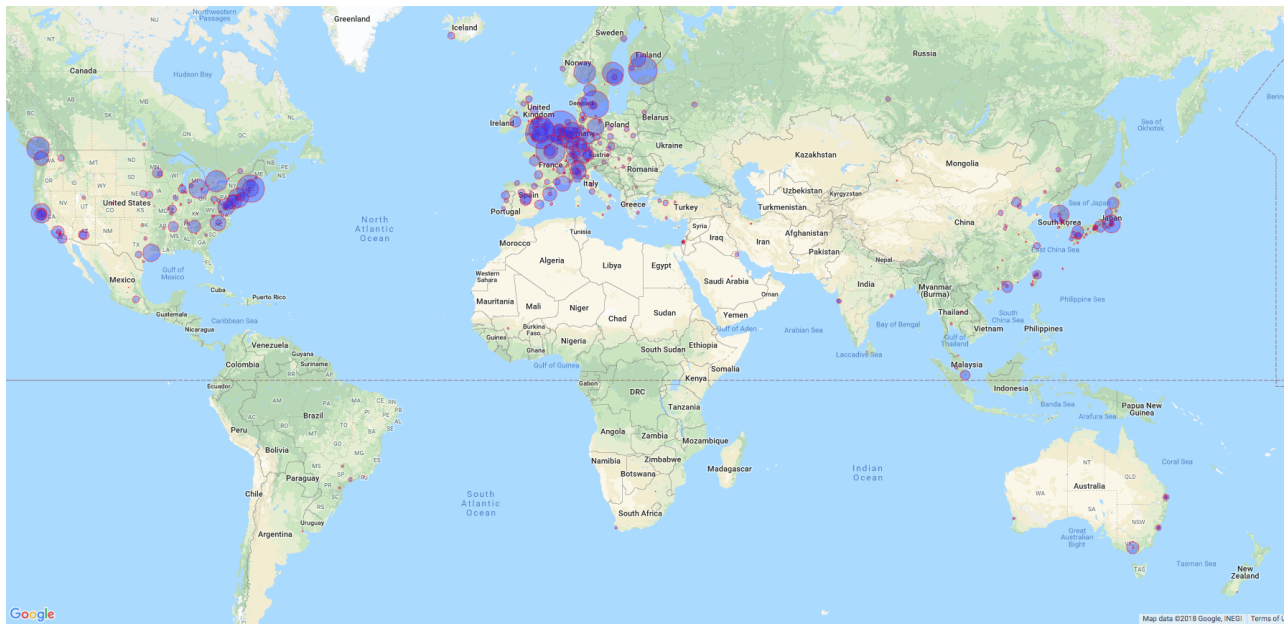


**Fig. 3.** Cumulative number of samples of the 15 most represented cancer types by ICD-O Topography codes in the arrayMap database, over a 10 year period (sample numbers in logarithmic scale)

rayMap 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 (figure 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.

A selection bias regarding cancer types also is apparent when comparing study representation (*arrayMap* and TCGA) with the respective incidences. While breast cancer cases represent 15.3% of all cancers (63), 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





**Fig. 4.** The map displays the geographic distribution (by corresponding author) of the 102225 genomic array, 36747 chromosomal CGH and 7212 whole genome/exome based cancer genome datasets. The numbers are derived from the 3103 publications registered in the Progenetix database.

sample representation; thyroid cancer 3.1% incidence with 0.16% (4.48%) samples; larynx carcinoma with 0.8% incidence and 0.05% of samples (1.11% TCGA).

Moreover, whereas some of the most studied cancers have low mortality rates such as breast with almost 90% survival after 5 years, special mention should be made for those entities underrepresented and with high mortality. For instance, pancreas cancers has 0.75% of samples in arraymap (1.63% TCGA) but 3.2% of all new cases with a 8.5% 5 years surviving 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 the 5 years 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 (64–66). 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 (67, 68). 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 on heritable contributions on cancer development.

Although, many studies relate the influence of geographic patterns incidences with environmental factors such as pollution levels, intensity of UV radiation or exposure to infectious agents (67), 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 outbreed reference populations (69). 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 followup questions or access to detailed study information or source data. However, this information can also be used as proxy to provide quantitative representation of study content with relation to geographic provenance, leading to some interesting observations.

The geographic origin mapping of the more than 3000 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 (1619) 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 socio-cultural practices, it is of paramount importance to include geo-location metadata in the assessment of molecular profiling. However, the real impact of factors correlated to geographic provenance can only be assessed with the availability of sufficient, representative data for a large range of geographies,

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

Pub.	Samples per Technique				(Sub-) Continent
	cCGH	aCGH	WES	WGS	
3	58	33	0	0	Africa
46	225	1444	303	288	Australia
465	6132	10736	1534	579	East Asia
28	564	1324	45	0	South Asia
31	720	209	0	0	Western Asia
1619	24070	47251	1912	1465	Europe
833	4680	39352	2888	936	North America
29	208	230	2	1	South America

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 the Europe, the United States, 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(70).

A more uniform map of the amount of genomic studies across the world is of further importance for clinical trials, where the 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 realised, a major obstacle in utilising the emerging data lies in the high degree of fragmentation and "siloeing" of the generated data. Genomics and associated metadata is frequently created as part of research studies with study specific consent(71) 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 (DAC) 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(72)) 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 boundaries. Formally established in 2014 and with and increasing participation of currently 499 organisational 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(73), standards and toolkits to enable secure and ethically responsible data sharing. These activities are based on the the work of different "work streams", which interact with existing "driver projects" in the iterative development, testing and implementation of protocols, standards and toolkits. The driver projects themselves - such as the "Beacon"(74) project or the "BRCA exchange"(75) 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 an guidelines for the effective sharing of genome-related data is a pre-requisite 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 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 lead 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 is 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 bioinformatics tools to create large surveys of existing genome studies now allows 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 promises "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 show-

ing that more than 50% of all reported genome variants in the Genome Aggregation Database (gnomAD) are based on European ancestry<sup>(76)</sup>.

Besides the well known impact of major socio-economic factors, efforts towards understanding of disparities in global cancer incidences and prognostic trajectories should also be directed with the characterisation 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 personalised 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 localised data silos and individual publications with, at best, highly fragmented but frequently non-existing access to genomic and associated metadata. Here, the Global Alliance for Genomics and Health provides a leading effort towards the better access to health related data, beyond individual studies and localised repositories, towards a global network of interacting standards and resources.

# Acknowledgements

MB wants to thank the Kavli Institute for Theoretical Physics (KITP), as guest of which parts of this manuscript were being written. Therefore, this research was supported in part by the National Science Foundation under Grant No. NSF PHY-1748958.

Additional thanks go to the current and former members of the Baudis group at UZH for their continuing efforts in making cancer genome data accessible.

# References

1. D Hanahan and RA Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5): 646–674, 2011.
2. J Lapointe, C Li, CP Giacomini, K Salari, S Huang, P Wang, M Ferrari, T Hernandez-Boussard, JD Brooks, and JR Pollack. Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. *Cancer Res*, 67(18):8504–8510, 2007.
3. DT Jones, N Jager, M Kool, T Zichner, B Hutter, M Sultan, YJ Cho, TJ Pugh, V Hovestadt, AM Stutz, T Rausch, HJ Warnatz, M Ryzhova, S Bender, D Sturm, S Pleier, H Cin, E Pfaff, L Sieber, A Wittmann, M Remke, H Witt, S Hutter, T Tzaridis, J Weischenfeldt, B Raeder, M Avci, V Amstislavskiy, M Zapotka, UD Weber, Q Wang, B Lasitschka, CC Bartholomae, M Schmidt, C von Kalle, V Ast, C Lawrenz, J Eils, R Kabbe, V Benes, P van Sluis, J Koster, R Volckmann, D Shih, MJ Betts, RB Russell, S Coco, GP Tonini, U Schuller, V Hans, N Graf, YJ Kim, C Monoranu, W Roggendorf, A Unterberg, C Herold-Mende, T Milde, AE Kulozik, A von Deimling, O Witt, E Maass, J Rossler, M Ebinger, MU Schuhmann, MC Fruhwald, M Hasselblatt, N Jabadó, S Rutkowski, AO von Bueren, D Williamson, SC Clifford, MG McCabe, VP Collins, S Wolf, S Wiemann, H Lehrach, B Brors, W Scheurlen, J Felsberg, G Reifenberger, PA Northcott, MD Taylor, M Meyerson, SL Pomeroy, ML Yaspo, JO Korbel, A Korshunov, R Eils, SM Pfister, and P Lichter. Dissecting the genomic complexity underlying medulloblastoma. *Nature*, 488(7409):100–105, 2012.
4. J Vandesompele, M Baudis, K De Preter, N Van Roy, P Ambros, N Bown, C Brinkschmidt, H Christiansen, V Combaret, M Lastowska, J Nicholson, A O'Meara, D Plantaz, R Stallings, B Brichard, C Van den Broecke, S De Bie, A De Paep, G Laureys, and F Speleman. Unequivocal delineation of clinicogenetic subgroups and development of a new model for improved outcome prediction in neuroblastoma. *J Clin Oncol*, 23(10):2280–2299, 2005.

5. D Malkin, FP Li, LC Strong, JF Fraumeni, CE Nelson, DH Kim, J Kassel, MA Gryka, FZ Bischoff, and MA Tainsky. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, 250(4985):1233–1238, 1990.
6. L Spirio, B Otterud, D Stauffer, H Lynch, P Lynch, P Watson, S Lanspa, T Smyrk, J Cavalieri, and L Howard. Linkage of a variant or attenuated form of adenomatous polyposis coli to the adenomatous polyposis coli (apc) locus. *Am J Hum Genet*, 51(1):92–100, 1992.
7. Y Miki, J Swensen, D Shattuck-Eidens, PA Futreal, K Harshman, S Tavtigian, Q Liu, C Cochran, LM Bennett, and W Ding. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182):66–71, 1994.
8. B Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, 2008.
9. 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.
10. El Heath, F Lynce, J Xiu, A Ellerbrock, SK Reddy, E Obeid, SV Liu, A Bollig-Fischer, D Separovic, and A Vanderwalde. Racial disparities in the molecular landscape of cancer. *Anti-cancer Res*, 38(4):2235–2240, 2018.
11. J. Kim, C.-L. Sun, B. Mailey, C. Prendergast, A. Artinyan, S. Bhatia, A. Pigazzi, and J. D. I. Ellenhorn. Race and ethnicity correlate with survival in patients with gastric adenocarcinoma. *Annals of Oncology*, 21(1):152–160, 2010. doi: 10.1093/annonc/mdp290.
12. 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.
13. Wensheng Zhang, Andrea Edwards, Erik K. Fleming, 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.
14. GN Calkins. Zur frage der entstehung maligner tumoren. *Science*, 40(1041):857–859, 1914.
15. PC Nowell and DA Hungerford. A minute chromosome in chronic granulocytic leukemia. *Science*, 132:1497, 1960.
16. JD Rowley. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243(5405): 290–293, 1973.
17. KD Zang and H Singer. Chromosomal constitution of meningiomas. *Nature*, 216(5110): 84–85, 1967.
18. F Mitelman, J Mark, PG Nilsson, H Dencker, C Norrby, and KG Tranberg. Chromosome banding pattern in human colonic polyps. *Hereditas*, 78(1):63–68, 1974.
19. Johansson B Mitelman F and Mertens F (Eds.). Mitelman database of chromosome aberrations and gene fusions in cancer, 2018.
20. A Kallioniemi, OP Kallioniemi, D Sudar, D Rutovitz, JW Gray, F Waldman, and D Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5182):818–821, 1992.
21. S Solinas-Toldo, S Lampel, S Stilgenbauer, J Nickolenko, A Benner, H Dohner, T Cremer, and P Lichter. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, 4(20):399–407, 1997.
22. PJ Campbell, PJ Stephens, ED Pleasance, S O'Meara, H Li, T Santarius, LA Stebbings, C Leroy, S Edkins, C Hardy, JW Teague, A Menzies, I Goodhead, DJ Turner, CM Clee, MA Quail, A Cox, C Brown, R Durbin, ME Hurler, PA Edwards, GR Bignell, MR Stratton, and PA Futreal. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, 40(6):722–729, 2008.
23. TJ Ley, ER Mardis, L Ding, B Fulton, MD McLellan, K Chen, D Dooling, BH Dunford-Shore, S McGrath, M Hickenbotham, L Cook, R Abbott, DE Larson, DC Koboldt, C Pohl, S Smith, A Hawkins, S Abbott, D Locke, LW Hillier, T Miner, L Fulton, V Magrini, T Wylie, J Glasscock, J Conyers, N Sander, X Shi, JR Osborne, P Minx, D Gordon, A Chinwalla, Y Zhao, RE Ries, JE Payton, P Westervelt, MH Tomasson, M Watson, J Baty, J Ivanovich, S Heath, WD Shannon, R Nagarajan, MJ Walter, DC Link, TA Graubert, JF DiPersio, and RK Wilson. Dna sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, 456(7218):66–72, 2008.
24. MM Le Beau, CA Westbrook, MO Diaz, and JD Rowley. c-src is consistently conserved in the chromosomal deletion (20q) observed in myeloid disorders. *Proc Natl Acad Sci U S A*, 82(19):6692–6696, 1985.
25. K Alitalo, M Schwab, CC Lin, HE Varmus, and JM Bishop. 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 U S A*, 6(80):1707–1711, 1983.
26. S Malcolm, P Barton, C Murphy, MA Ferguson-Smith, DL Bentley, and TH Rabbitts. 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 U S A*, 79(16):4957–4961, 1982.
27. M Schwab, J Ellison, M Busch, W Rosenau, HE Varmus, and JM Bishop. Enhanced expression of the human gene n-myc consequent to amplification of dna may contribute to malignant progression of neuroblastoma. *Proc Natl Acad Sci U S A*, 15(81):4940–4944, 1984.
28. P Lichter, T Cremer, J Borden, L Manuelidis, and DC Ward. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant dna libraries. *Hum Genet*, 80(3):224–234, 1988.
29. DC Tkachuk, CA Westbrook, M Andreeff, TA Donlon, ML Cleary, K Suryanarayan, M Homge, A Redner, J Gray, and D Pinkel. Detection of bcr-abl fusion in chronic myelogenous leukemia by in situ hybridization. *Science*, 250(4980):559–562, 1990.
30. MR Speicher, S Gwyn Ballard, and DC Ward. Karyotyping human chromosomes by combi-



- natorial multi-fluor fish. *Nat Genet*, 12(4):368–375, 1996.
31. T Veldman, C Vignon, E Schrock, JD Rowley, and T Ried. Hidden chromosome abnormalities in haematological malignancies detected by multicolour spectral karyotyping. *Nat Genet*, 15(4):406–410, 1997.
  32. S Joos, H Scherthan, MR Speicher, J Schlegel, T Cremer, and P Lichter. Detection of amplified dna sequences by reverse chromosome painting using genomic tumor dna as probe. *Hum Genet*, 90(6):584–589, 1993.
  33. MR Speicher, S du Manoir, E Schrock, H Holtgreve-Grez, B Schoell, C Lengauer, T Cremer, and T Ried. Molecular cytogenetic analysis of formalin-fixed, paraffin-embedded solid tumors by comparative genomic hybridization after universal dna-amplification. *Hum Mol Genet*, 11(2):1907–1914, 1993.
  34. J Isola, S DeVries, L Chu, S Ghazvini, and F Waldman. Analysis of changes in dna sequence copy number by comparative genomic hybridization in archival paraffin-embedded tumor samples. *Am J Pathol*, 6(145):1301–1308, 1994.
  35. M Bentz, A Plesch, S Stiltgenbauer, H Dohner, and P Lichter. Minimal sizes of deletions detected by comparative genomic hybridization. *Genes Chromosomes Cancer*, 2(21):172–175, 1998.
  36. O Monni, H Joensuu, K Franssila, J Klefstrom, K Alitalo, and S Knuutila. Bcl2 overexpression associated with chromosomal amplification in diffuse large b-cell lymphoma. *Blood*, 90(3):1168–1174, 1997.
  37. CA Werner, H Dohner, S Joos, LH Trumper, M Baudis, TF Barth, G Ott, P Moller, P Lichter, and M Bentz. High-level dna amplifications are common genetic aberrations in b-cell neoplasms. *Am J Pathol*, 151(2):335–342, 1997.
  38. S Knuutila, AM Bjorkqvist, K Autio, M Tarkkanen, M Wolf, O Monni, J Szymanska, ML Larremendy, J Tapper, H Pere, W El-Rifai, S Hemmer, VM Wasenius, V Vidgren, and Y Zhu. Dna copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. *Am J Pathol*, 152(5):1107–1123, 1998.
  39. RG Weber, J Bostrom, M Wolter, M Baudis, VP Collins, G Reifenberger, and P Lichter. Analysis of genomic alterations in benign, atypical, and anaplastic meningiomas: toward a genetic model of meningioma progression. *Proc Natl Acad Sci U S A*, 26(94):14719–14724, 1997.
  40. D Pinkel, R Segreaves, D Sudar, S Clark, I Poole, D Kowbel, C Collins, WL Kuo, C Chen, Y Zhai, SH Dairkee, BM Ljung, JW Gray, and DG Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, 2(20):207–211, 1998.
  41. JR Pollack, CM Perou, AA Alizadeh, MB Eisen, A Pergamenschikov, CF Williams, SS Jeffrey, D Botstein, and PO Brown. Genome-wide analysis of dna copy-number changes using cdna microarrays. *Nat Genet*, 1(23):41–46, 1999.
  42. DG Wang, JB Fan, CJ Siao, A Berno, P Young, R Sapolsky, G Ghandour, N Perkins, E Winchester, J Spencer, L Kruglyak, L Stein, L Hsie, T Topaloglou, E Hubbell, E Robinson, M Mittmann, MS Morris, N Shen, D Kilburn, J Rioux, C Nusbaum, S Rozen, TJ Hudson, R Lipshutz, M Chee, and ES Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, 1998.
  43. X Zhao, C Li, JG Paez, K Chin, PA Jänne, TH Chen, L Girard, J Minna, D Christiani, C Leo, JW Gray, WR Sellers, and M Meyerson. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*, 64(9):3060–3071, 2004.
  44. Bernd Timmermann, Martin Kerick, Christina Roehr, Axel Fischer, Melanie Isau, Stefan T Boerno, Andrea Wunderlich, Christian Barmeyer, Petra Seemann, Jana Koenig, Michael Lappe, Andreas W Kuss, Masoud Garshasbi, Lars Bertram, Kathrin Trappe, Martin Werber, Bernhard G Herrmann, Kurt Zatloukal, Hans Lehrach, and Michal R Schweiger. Somatic mutation profiles of msi and mss colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PLoS ONE*, 5(12):e15661, 2010.
  45. M. Meyerson, S. Gabriel, and G. Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, 11(10):685–696, 2010.
  46. J de Ligt, PM Boone, R Pfundt, LE Visser, T Richmond, J Geoghegan, K O’Moore, N de Leeuw, C Shaw, HG Brunner, JR Lupski, JA Veltman, and JY Hehir-Kwa. Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum Mutat*, 34(10):1439–1448, 2013.
  47. W Wang, W Sun, JJ Crowley, and JP Szatkiewicz. Allele-specific copy-number discovery from whole-genome and whole-exome sequencing. *Nucleic Acids Res*, 43(14):e90, 2015.
  48. Costas S Iliopoulos and Solon P Pissis. Algorithms for next-generation sequencing data. *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data*, pages 251–280, 2013.
  49. DG Holland, A Burleigh, A Git, MA Goldgraben, PA Perez-Mancera, SF Chin, A Hurtado, A Bruna, HR Ali, W Greenwood, MJ Dunning, S Samarajiva, S Menon, OM Rueda, AG Lynch, S McKinney, IO Ellis, CJ Eaves, JS Carroll, C Curtis, S Aparicio, and C Caldas. Znf703 is a common luminal b breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. *EMBO Mol Med*, 3(3):167–180, 2011.
  50. PA Northcott, DJ Shih, J Peacock, L Garzia, AS Morrissy, T Zichner, AM Stutz, A Korshunov, J Reimand, SE Schumacher, R Beroukhi, DW Ellison, CR Marshall, AC Lionel, S Mack, A Dubuc, Y Yao, V Ramaswamy, B Luu, A Rolider, FM Cavalli, X Wang, M Remke, X Wu, RY Chiu, A Chu, E Chuah, RD Corbett, GR Hoad, SD Jackman, Y Li, A Lo, KL Mungall, KM Nip, QJ Qian, AG Raymond, NT Thiessen, RJ Varhol, I Birol, RA Moore, AJ Mungall, R Holt, D Kawauchi, MF Roussel, M Kool, DT Jones, H Witt, A Fernandez-L, AM Kenney, RJ Wechsler-Reya, P Dirks, T Aviv, WA Grajkowska, M Perek-Polnik, CC Haberler, O Delattre, SS Reynaud, FF Doz, SS Pernet-Fattet, BK Cho, SK Kim, KC Wang, W Scheurlen, CG Eberhart, M Fevre-Montange, A Jouveit, IF Pollack, X Fan, KM Muraszko, GY Gillespie, C Di Rocco, L Massimi, EM Michiels, NK Kloosterhof, PJ French, JM Kros, JM Olson, RG Ellenbogen, K Zitterbart, L Kren, RC Thompson, MK Cooper, B Lach, RE McLendon, DD Bigner, A Fontebasso, S Albrecht, N Jabado, JC Lindsey, S Bailey, N Gupta, WA Weiss, L Bognar, A Klekner, TE Van Meter, T Kumabe, T Tominaga, SK Elbabaa, JR Leonard, JB Rubin, LM Liau, EG Van Meir, M Fouladi, H Nakamura, G Cinalli, M Garami, P Hauser, AG Saad, A Iolascon, S Jung, CG Carloti, R Vibhakara, YS Ra, S Robinson, M Zollo, CC Faria, JA Chan, ML Levy, PH Sorensen, M Meyerson, SL Pomeroy, YJ Cho, GD Bader, U Tabori, CE Hawkins, E Bouffet, SW Scherer, JT Rutka, D Malkin, SC Clifford, SJ Jones, JO Korbel, SM Pfister, MA Marra, and MD Taylor. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature*, 488(7409):49–56, 2012.
  51. M Baudis and ML Cleary. Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics*, 17(12):1228–1229, 2001.
  52. 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.
  53. H Cai, N Kumar, N Ai, S Gupta, P Rath, and M Baudis. Progenetix: 12 years of oncogenomic data curation. *Nucleic Acids Res*, 42(Database issue):D1055–62, 2014.
  54. CA Werner, H Dohner, S Joos, LH Trumper, M Baudis, TF Barth, G Ott, P Moller, P Lichter, and M Bentz. High-level dna amplifications are common genetic aberrations in b-cell neoplasms. *Am J Pathol*, 151(2):335–342, 1997.
  55. R Beroukhi, CH Mermel, D Porter, G Wei, S Raychaudhuri, J Donovan, J Barretina, JS Boehm, J Dobson, M Urashima, KT Mc Henry, RM Pinchback, AH Ligon, YJ Cho, L Haery, H Greulich, M Reich, W Winkler, MS Lawrence, BA Weir, KE Tanaka, DY Chiang, AJ Bass, A Loo, C Hoffman, J Prensner, T Liefeld, Q Gao, D Yecies, S Signoretti, E Maher, FJ Kaye, H Sasaki, JE Tepper, JA Fletcher, J Tabernero, J Baselga, MS Tsao, F Demicheli, MA Rubin, PA Janne, MJ Daly, C Nucera, RL Levine, BL Ebert, S Gabriel, AK Rustgi, CR Antonescu, M Ladanyi, A Letai, LA Garraway, M Loda, DG Beer, LD True, A Okamoto, SL Pomeroy, S Singer, TR Golub, ES Lander, G Getz, WR Sellers, and M Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, 2010.
  56. R Beroukhi, G Getz, L Nghiemphu, J Barretina, T Hsueh, D Linhart, I Vivanco, JC Lee, JH Huang, S Alexander, J Du, T Kau, RK Thomas, K Shah, H Soto, S Perner, J Prensner, RM Debiasi, F Demicheli, C Hattori, MA Rubin, LA Garraway, SF Nelson, L Liau, PS Mischel, TF Cloughesy, M Meyerson, TA Golub, ES Lander, IK Mellinghoff, and WR Sellers. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A*, 104(50):20007–20012, 2007.
  57. T Sjöblom, S Jones, LD Wood, DW Parsons, J Lin, TD Barber, D Mandelker, RJ Leary, J Ptak, N Silliman, S Szabo, P Buckhaults, C Farrell, P Meeh, SD Markowitz, J Willis, D Dawson, JK Willson, AF Gazdar, J Hartigan, L Wu, C Liu, G Parmigiani, BH Park, KE Bachman, N Papadopoulos, B Vogelstein, KW Kinzler, and VE Velculescu. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–274, 2006.
  58. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.
  59. TJ Hudson and International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.
  60. EMBL-EBI. Eml-ebi arrayexpress. Accessed: 2017-06-30.
  61. NCBI. Ncbi gene expression omnibus. [www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/). Accessed: 2017-06-30.
  62. A Fritz, C Percy, A Jack, LH Sobin, and MD Parkin, editors. *International Classification of Diseases for Oncology (ICD-O), Third Edition*. World Health Organization, Geneva, 2000.
  63. National Cancer Institute. Surveillance, Epidemiology, and End Results Program. <https://seer.cancer.gov/statfacts/>, 2018.
  64. PG Casali. Rare cancers: work in progress in europe. *Ann Oncol*, 25(4):914, 2014.
  65. G Bevilacqua, F Bosman, T Dasse, H Höfler, A Janin, R Langer, D Larisimont, MM Morente, P Riegman, P Schirmacher, G Stanta, K Zatloukal, E Caboux, and P Hainaut. The role of the pathologist in tissue banking: European consensus expert group report. *Virchows Arch*, 456(4):449–454, 2010.
  66. ES Dove. Biobanks, data sharing, and the drive for a global privacy governance framework. *J Law Med Ethics*, 43(4):675–689, 2015.
  67. 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.
  68. RL Siegel, KD Miller, and A Jemal. Cancer statistics, 2017. *CA Cancer J Clin*, 67(1):7–30, 2017.
  69. CI Szabo and MC King. Population genetics of brca1 and brca2. *Am J Hum Genet*, 60(5):1013–1020, 1997.
  70. M Helmy, M Awad, and KA Mosa. Limited resources of genome sequencing in developing countries: Challenges and solutions. *Appl Transl Genom*, 9:15–19, 2016.
  71. SO Dyke, AA Philippakis, J Rambla De Argila, DN Paltoo, ES Luetkemeyer, BM Knoppers, AJ Brookes, JD Spalding, M Thompson, M Roos, KM Boycott, M Brudno, M Hurles, HL Rehm, A Matern, M Fiume, and ST Sherry. Consent codes: Upholding standard data use conditions. *PLoS Genet*, 12(1):e1005772, 2016.
  72. M Lawler, LL Siu, HL Rehm, SJ Chanock, G Alterovitz, J Burn, F Calvo, D Lacombe, BT Teh, KN North, CL Sawyers, Clinical Working Group of the Global Alliance for Genomics and (GA4GH Health). All the world’s a stage: Facilitating discovery science and improved cancer care through the global alliance for genomics and health. *Cancer Discov*, 5(11):1133–1136, 2015.
  73. 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: 2018-05-07.
  74. ELIXIR. Elixir beacon. <https://www.elixir-europe.org/about-us/implementation-studies/beacons>. Accessed: 2018-05-07.
  75. BRCA Exchange. Brca exchange of the global alliance for genomics and health. <http://brcaexchange.org>. Accessed: 2018-05-07.
  76. Alice B Popejoy, Deborah I Ritter, Kristy Crooks, Erin Currey, Stephanie M Fullerton, Lucia A Hindorf, Barbara Koenig, Erin M Ramos, Elena P Sorokin, Hannah Wand, Matthew W Wright, James Zou, Christopher R Gignoux, Vence L Bonham, Sharon E Plon, and Carlos D Bustamante. The clinical imperative for inclusivity: Race, ethnicity, and ancestry (rea) in genomics. *bioRxiv*, 2018.