Laboratory Investigation (2019) 99:118–127
https://doi.org/10.1038/s41374-018-0125-5

ARTICLE

Trends in the characteristics of human functional genomic data on the gene expression omnibus, 2001–2017

Daniel D. Liu[1] · Lanjing Zhang[2,3,4,5]

Received: 8 February 2018 / Revised: 25 July 2018 / Accepted: 15 August 2018 / Published online: 11 September 2018

Abstract

The gene expression omnibus (GEO) is the world's largest public repository of functional genomic data. Despite its broad use in secondary genomic analyses, the temporal trends in the characteristics of genomic data on GEO, including experimental procedures, geographic origin, funder(s), and related disease, have not been examined. We identified 75,376 Series deposited to the GEO during 2001–2017 and built a database of all human genomic data (39,076 Series, 51.8% of all Series). Using the associated publications, we obtained funding information and identified the related disease area. Of the Series with classified disease areas, the two most common were cancer ($n = 12{,}688$, 32.5%) and immunologic diseases ($n = 2{,}393$, 6.1%), while the percentages of all other disease areas were below 5%, including neurological diseases ($n = 1733$, 4.4%), infectious diseases ($n = 1225$, 3.1%), diabetes ($n = 828$, 2.1%), and cardiovascular diseases ($n = 299$, 0.8%). In recent years, there has been a significant increase in the use of high-throughput sequencing (HTS), protein array and multiple-platform technologies, as well as in the proportion of North American deposits. Compared to those from other regions, North American deposits appeared to lead the shift from array-based to HTS technologies (odds ratio [OR], 95% confidence intervals [CI] = 3.39, 3.23–3.55, $P = 9.40\text{E}{-323}$), and were less likely to focus on a major disease area (OR = 0.64, 95% CI: 0.61–0.67, $P = 5.02\text{E}{-107}$), suggesting a greater emphasis on basic science in North America. Furthermore, the Series utilizing HTS were less likely to be disease-classified compared to other technologies (OR = 0.39, 95% CI: 0.37–0.41, $P = 1.00\text{E}{-322}$), suggesting a preferential use or adoption of HTS in basic science settings. Finally, funding from the NHGRI, NCI, NIEHS, and NCCR resulted in a higher number of GEO Series per grant than other NIH institutes, demonstrating different preferences on genomic studies among awardees of NIH institutes. Our findings demonstrate geographic, technological, and funding disparities in the trends of GEO deposit characteristics.

Introduction

The gene expression omnibus (GEO) is the world's largest public repository of functional genomic data, founded and run by the US-based National Center for Biotechnology Information (NCBI) within the National Library of Medicine at the National Institutes of Health (NIH) [1]. Along with its European counterpart ArrayExpress [2], such repositories are central towards fostering reproducibility and open access in genomic research [3].

GEO data are classified into four entity types: Platform (GPL), Sample (GSM), Series (GSE), and DataSet (GDS) [1]. Platform (GPL) records detail the specific technology or technologies used to obtain data of a given sample. Sample (GSM) records describe the experimental output of one individual sample. Series (GSE) records consist of a group of related Samples within an experiment. Finally, DataSet (GDS) records are the Series that have been curated by

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41374-018-0125-5) contains supplementary material, which is available to authorized users.

✉ Lanjing Zhang
lanjing.zhang@rutgers.edu
ljzhang@hotmail.com

[1] Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

[2] Department of Pathology, University Medical Center of Princeton, Plainsboro, NJ 08536, USA

[3] Department of Biological Sciences, Rutgers University, Newark, NJ 07102, USA

[4] Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903, USA

[5] Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ 08854, USA

GEO staff, normalized to be biologically and statistically comparable.

Buried within the metadata of GEO deposits, however, lie broader trends in the research ecosystem. Open-access genomic databases on human samples are critical for future advances in oncology and medicine, and have been expanded significantly in the past decade [4–10]. However, to date, there have been no in-depth analyses of the trends in functional genomic data on GEO or ArrayExpress, despite their growing importance and volume. Such information could prove especially useful for the research on genomic medicine [11], public health [12, 13], and science funding and policy [14].

Here, we developed a database of human GSE alongside their associated metadata, and identified the temporal trends in genomic data growth on GEO. Only some of this metadata was readily available from the GEO browser; the disease-of-interest was extracted from experiment summaries, and funding data were extracted from the associated publications. Probing this database yielded several new insights on the technology, geographic origin, and research focus of the functional genomic studies. Most prominently, we observed a rapid adoption of high-throughput sequencing (HTS) in North America, alongside a shift toward basic research in human.

Materials and methods

Metadata extraction

We identified and included human GEO series using the organism keyword of *Homo sapiens* without any other search criteria in July 2017, and again in January 2018 for updates. Metadata on all human GEO Series (GSE) were downloaded from the GEO repository browser, including accession codes, title, Series type, release date, and associated curated GDS. Geographic origin (i.e., the corresponding author's affiliation on the record) and experimental summaries were extracted from each Series' accession display page using a custom web scraper. For Series with one or more associated publications, further metadata were extracted from MEDLINE, a bibliographic database indexed by the National Library of Medicine. We extracted the grant numbers under the GR field, and the medical subject headings from the MH field. Only the Series uploaded on or before 31 December 2017 were included in the analyses.

Data curation

From the raw metadata, certain fields were extracted to facilitate analyses. The Series type indicates both the general application (e.g., expression profiling or SNP genotyping) and the technology used (e.g., array or HTS). Due to the large number of such combinations, we separated the application and technology for individual analysis.

We classified each Series into one of the six broad disease areas using a keywords-based classification strategy: cancer, cardiovascular diseases, diabetes, immunologic, infectious diseases, and neurologic diseases. Briefly, we scanned each Series' summary for keywords relating to each disease classification (Supplementary Table 1), and categorized it into the one with the greatest number of keyword hits. Those with no keyword hits were categorized as "unclassified."

From the grant numbers, we parsed out the specific National Institutes of Health (NIH) institute(s) funding each grant, or listed down "other" for non-NIH institutes. During the data analysis, if a Series was funded by more than one NIH institute or center, each was counted once. If a Series was funded by two grants from the same institute, that institute was counted twice.

Statistical analysis

Statistical analyses including Fischer exact test were performed using MATLAB (Version R2017a March 2017, MathWorks). The Joinpoint Regression Program (Version 4.5.0.1. June 2017, Statistical Research and Applications Branch, National Cancer Institute, Bethesda, MD, USA) was used to analyze the trends in the number of deposited Series per annum and subgroup trend-analyses, from which annual percent change (APC) values were computed [15]. The model selections were based on permutation tests in which log transformation was conducted, an overall P value < 0.05 was considered as significant, and the number of randomly permuted data sets was 4499. Up to two joinpoints were allowed. All P values were two-sided.

Results

Of the 75,376 Series deposited on GEO between 2001 and 2017, a total of 39,076 (51.8%) were human samples. Raw data for the human Series are summarized in Table 1. Fig. 1 shows the descriptive statistics of the Series by geographic origin, disease classification, genomic application, and technology. A slight majority of Series (54%) originated from North America, followed by Europe (28%) and Asia (15%) (Fig. 1a). Around 48% of Series could be classified to one of six major disease-categories: in descending order, cancer (30%), immunologic diseases (9%), neurologic diseases (4%), infectious diseases (3%), diabetes (2%), and

Table 1 Trends in the characteristics of the functional genomic data deposited in the gene expression omnibus (GEO) from 2001–2017

			Region				Technology		Disease area						
Year	Series	GDS	Asia	Europe	N. Am.	Other	Array	HTS	Cancer	Cardio	Diabetes	Immune	Infectious	Neuro	Unclass.
2001	4	0	0	1	3	0	2	0	3	0	0	0	0	0	1
2002	38	8	0	3	35	0	28	0	6	0	13	0	1	1	17
2003	186	50	3	24	153	6	178	0	16	4	5	16	12	4	129
2004	268	121	10	58	192	8	253	0	57	9	8	21	12	9	152
2005	468	160	30	100	325	13	408	0	144	7	10	46	23	24	214
2006	612	193	47	177	376	12	538	0	205	7	13	68	22	32	265
2007	860	184	69	239	522	30	706	0	304	3	20	113	26	45	349
2008	1169	120	129	359	641	40	951	16	466	7	28	157	32	45	434
2009	1599	80	170	509	860	60	1252	39	622	5	35	202	32	73	630
2010	2026	84	273	648	1026	79	1538	115	719	18	52	252	62	97	826
2011	2648	200	375	789	1366	118	1990	219	1019	16	50	322	60	115	1066
2012	3088	134	485	1047	1432	124	2215	334	1191	32	66	330	103	148	1218
2013	3514	117	592	1136	1670	116	2438	480	1325	18	80	363	90	181	1457
2014	4002	112	742	1322	1758	180	2522	810	1417	40	85	436	120	206	1698
2015	4551	30	784	1550	2045	172	2713	1037	1618	38	105	486	152	226	1926
2016	6147	2	970	1401	3586	190	2409	2905	1713	61	145	517	150	289	3272
2017	7896	0	1156	1557	4957	226	2325	4379	1821	40	130	340	187	316	5062
Total	39,076	1595	5835	10,920	20,947	1374	22,466	10,334	12,688	299	828	2393	1225	1733	19910

Note: *N. Am.* North America; *HTS* high-throughput sequencing; *cardio* cardiovascular diseases; *immune* immunologic diseases; *neuro* neurologic diseases; *unclass.* unclassified. All deposits not fit in any of the six disease areas were categorized as unclassified

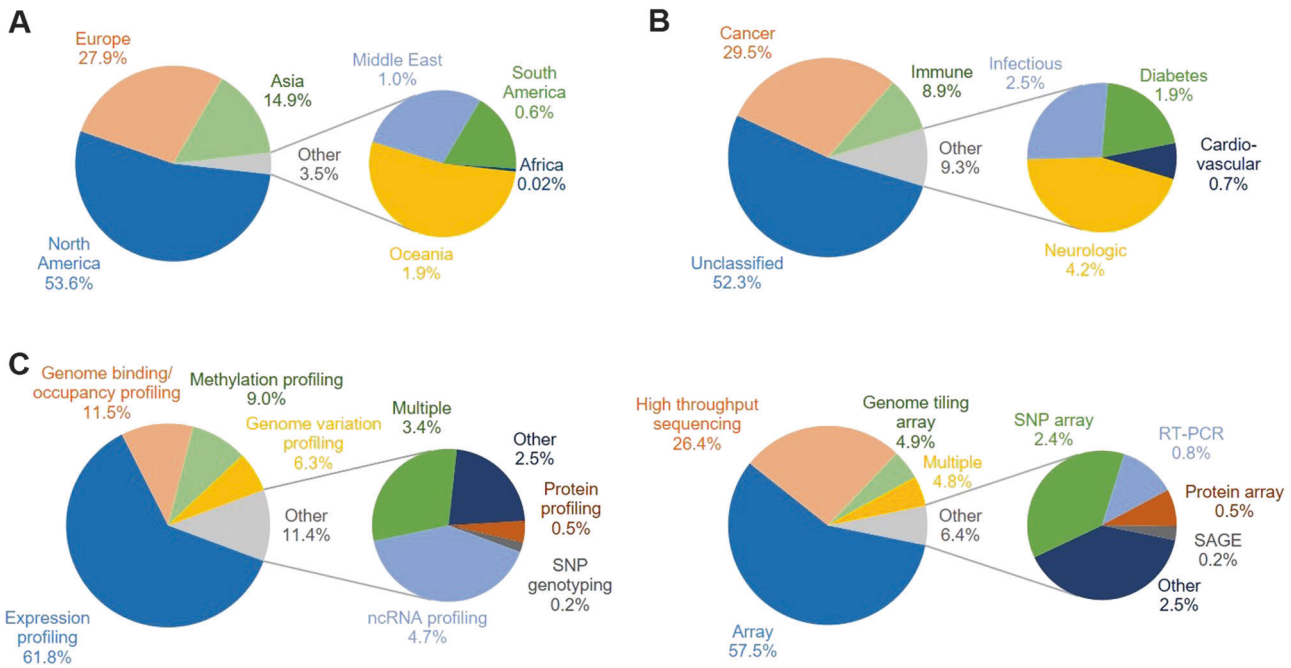

Fig. 1 Summary statistics. **a** Geographic origin of submitted GSE. **b** Area of study for submitted GSE. **c** Platform usage, separated into application (left) and technology (right)

cardiovascular diseases (1%) (Fig. 1b). The remaining "unclassified" Series consisted of mostly basic science studies, and some less prevalent diseases. Genomic application was dominated by expression profiling (62%), (Fig. 1c). The majority of the Series were collected using array technologies (58%) or HTS (26%) (Fig. 1d).

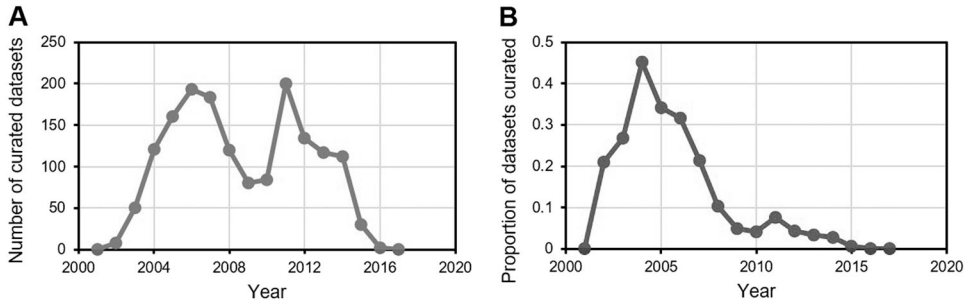

Fig. 2 Curated datasets. **a** Absolute number of curated DataSets (GDS). Year indicates the submission date of the associated Series (**b**) Proportion of each years' submitted GSE that have been curated into GDS

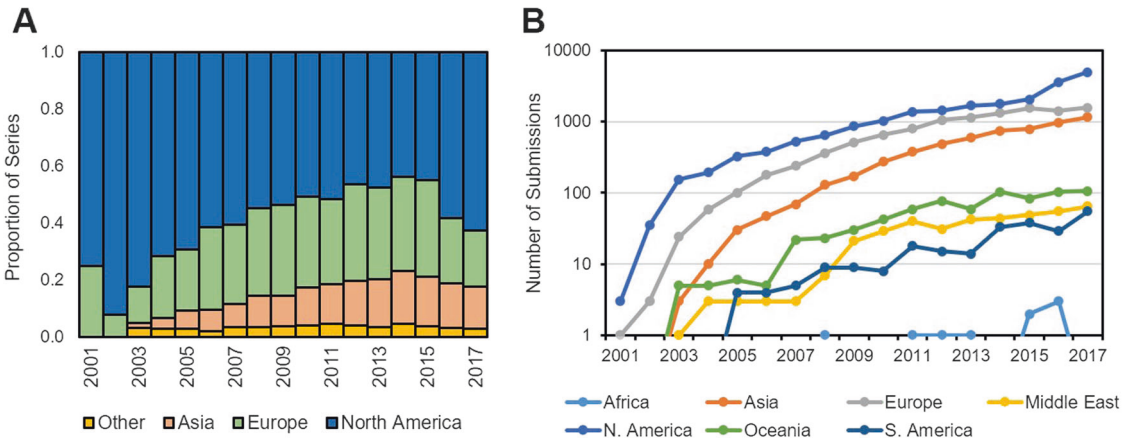

Fig. 3 Geographic distribution. **a** Proportion and **b** absolute number of datasets originating from specific geographical regions. Absolute numbers for 2017 are projected off the first 3 months of the year

We next sought to discover trends in these data over time. In regards to the number of Series deposited per year, we identified two segments of growth (one joinpoint), namely 2001–2009 (APC = 43.6, $P < 0.001$) and 2009–2017 (APC = 20.3, $P < 0.001$). Sharp fluctuations were found in the number of DataSets (GDS) curated from each year (Fig. 2). GDS curation grew rapidly from 2001 to 2006, when it peaked at 193, but following this period, a very low number of Series were curated from 2008 to 2010. In 2011 there was a sudden jump up to 200 GDS, but the number has since dropped to zero.

There were also trends in the geographic origin of Series (Fig. 3). When GEO was launched, a vast majority of the submitted Series originated from North America. With each passing year, however, Europe and Asia represented an increasingly large proportion of submitted Series. This trend took a dramatic turn in 2015, after which the proportion of North American Series sharply increased (Fig. 3a). Analysis of the raw number of Series per year shows that European deposits have plateaued around 2012, with other regions still steadily growing (Fig. 3b).

Given the rapidly evolving nature of genomics, it is perhaps unsurprising that there were changing trends in the genomic technologies used for producing the deposited human genomic data. While array-based technologies initially predominated, HTS rapidly overtook it in 2016 (Fig. 4a). The number of HTS Series deposited per year has been exponentially increasing (APC = 79 for 2009–2017, $P < 0.001$), while arrays have nearly plateaued in recent years (APC = 3.4 for 2011–2017, $P = 0.07$) (Fig. 4b, Supplementary Table S2). There has also been a sustained increase in the number of Series using "other" technologies (APC = 59 for 2001–2017, $P < 0.001$), possibly reflecting the growing number of emerging functional genomic techniques. Interestingly, Series originated from North America were 3 times more likely to use HTS technology compared to those from other regions (OR = 3.39), a gap that dramatically widened after 2015 ($OR_{2017} = 5.52$) (Fig. 4c, Table 2).

We next investigated trends in the Series' disease-of-interest over time. The proportion of Series that could be classified to one of the six major disease-categories increased steadily from 2003 to around 2008, after which it remained steady at around 60% (Fig. 5a, b). However, starting in 2015, the proportion of Series related to major disease area dropped sharply, down to 36% in 2017. This reflects an increase in "unclassified" Series focusing on basic science and less prevalent diseases. Nevertheless, all six disease classifications still saw a steady growth in the number of Series per year (Supplementary Table S3). The

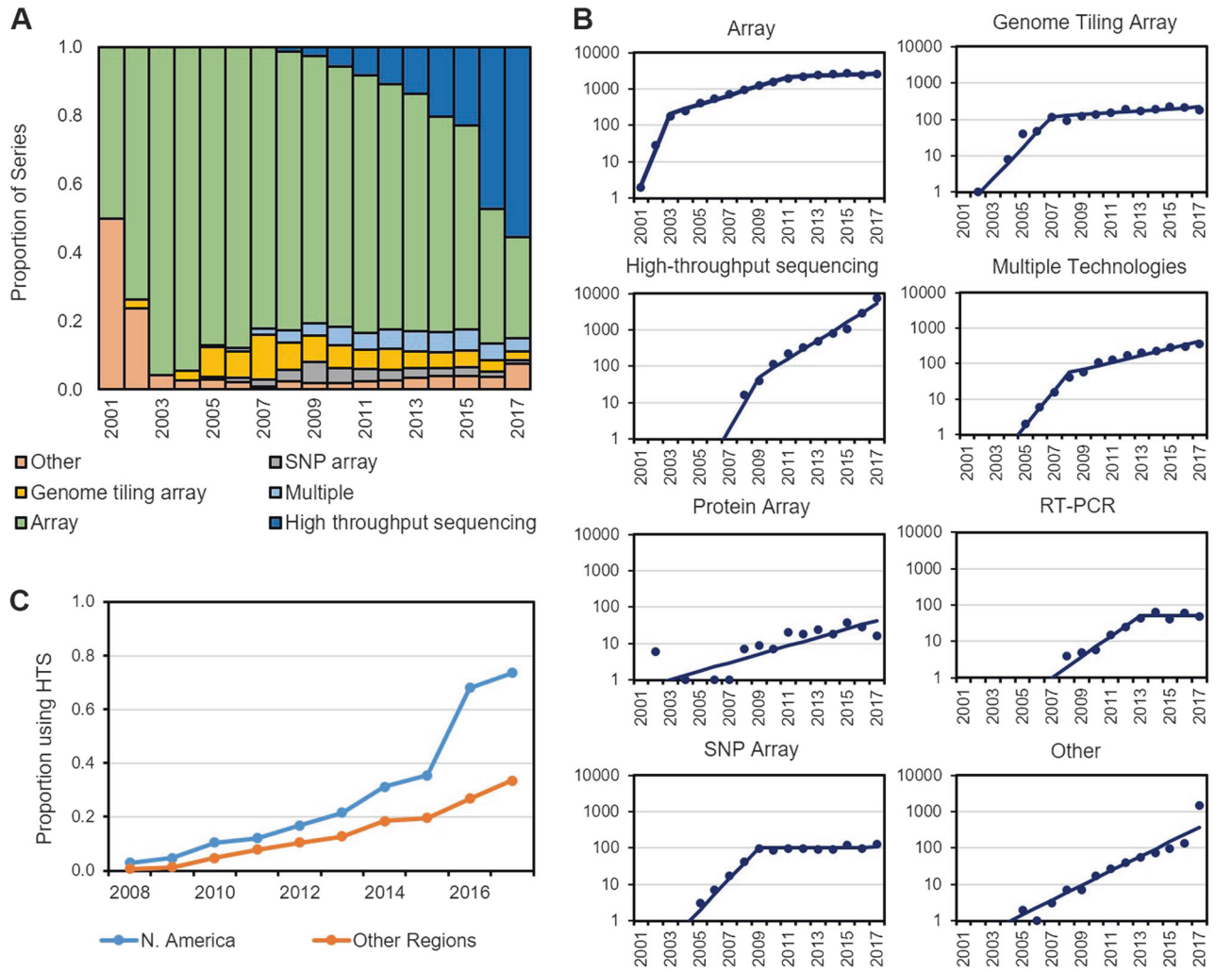

Fig. 4 Genomic platform usage. **a** Proportion of datasets using certain technologies, by year. **b** Joint point analysis of each technology's absolute growth. **c** Proportion of datasets using high-throughput sequencing (HTS) in North America vs. other regions from 2008 onwards

decreasing proportion of disease-classified Series was due almost entirely to those of North America, which dropped from 59% disease-classified in 2015 to just 25% in 2017, while there was no change for the rest of the world (Fig. 5c, Table 3). Importantly, Series utilizing HTS were significantly less likely to be disease-classified compared to other technologies (OR = 0.39), suggesting a preferential use or adoption of HTS in basic science settings (Fig. 5d, Table 4).

Finally, we assessed trends in the funding sources of Series with associated publication(s) indexed in the MEDLINE. Funding information could only be extracted and analyzed for Series with associated publications, accounting for ~68% of all Series. Of the grants with associated publications indexed in MEDLINE, the large majority (86%) were funded by the U.S. NIH. The NIH institutes funding the greatest proportion of Series were, in descending order, the National Cancer Institute (NCI, 33%), National Institute on Aging (NIA, 11%), National Institute of General Medical Sciences (NIGMS, 7.7%), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK, 6.7%), and National Heart, Lung, and Blood Institute (NHLBI, 6.6%). There were no significant trends in funding sources over time (Fig. 5e). However, simply assessing the proportion of Series funded by a particular agency can be misleading, as larger agencies can naturally fund more studies. To address this, we normalized the number of Series funded by each NIH institute to the total number of grants funded by that institute, giving the proportion of grants that result in a GEO Series. The overall NIH proportion was 0.063, or nearly one Series produced per 16 grants. Five institutes were above this level: unsurprisingly, the National Human Genome Research Institute (NHGRI, 0.49), followed by the NIA (0.19), NCI (0.18), National Center for Research Resources (NCRR, 0.12), and National Institute of Environmental Health Sciences (NIEHS, 0.085) (Fig. 5f). The NIH was not more likely to fund disease-classified studies compared with non-US

Table 2 Association between human GEO deposits' technology used (high-throughput sequencing technology vs. other methods) and their corresponding geographic origin (North America vs. other regions) from 2008 to 2017

	North America		Other regions				
Year	HTS	Other method	HTS	Other method	OR	95% CI	*P* value
2008	18	623	3	525	5.06	1.48–17.26	3.59E−03
2009	40	820	9	730	3.96	1.91–8.21	5.70E−05
2010	106	920	46	954	2.39	1.67–3.42	8.07E−07
2011	165	1201	100	1182	1.62	1.25–2.11	2.74E−04
2012	240	1192	173	1483	1.73	1.40–2.13	3.27E−07
2013	360	1310	233	1611	1.90	1.59–2.28	2.16E−12
2014	549	1209	413	1830	2.01	1.74–2.33	7.93E−21
2015	726	1319	488	2017	2.27	1.99–2.60	7.31E−34
2016	2436	1150	684	1877	5.81	5.20–6.50	1.31E−229
2017	3643	1313	983	1955	5.52	5.00–6.09	6.33E−270
Total	8283 (22.6%)	11,057 (30.2%)	3132 (8.5%)	14,164 (38.7%)	3.39	3.23–3.55	9.40E−323

Note: Odds ratio (OR) represents the odds that a GEO deposit using HTS originated from North America. No deposits using HTS were identified in the GEO prior to 2008. *HTS* high-throughput sequencing; *CI* confidence intervals

agencies (OR = 1.02, $P = 0.677$) (Supplementary Table S4).

Discussion

Since its inception in 2001, the GEO has become a mainstay of molecular biology research [1]. Its exponential growth reflects an evolving research environment where HTS technologies are increasingly used in human genomic studies. GEO metadata thus present a valuable resource in analyzing trends in the research ecosystem. This study, to our best knowledge, represents the first in-depth study of human GEO Series, encompassing geography, disease of interest, funding sources, genomic application, and technology. The summary database curated here is powerful because it not only allows for analysis of descriptive statistics and trends, but also correlations that offer clues as to the origin of specific trends.

Curated DataSets (GDS) are very valuable tools for researchers. They are normalized to be biologically comparable, and are compatible with a suite of data display and analysis tools offered by GEO. Thus, the sharp decline in GDS records in recent years may be troublesome for high-quality, secondary genomic analyses. However, due to the increasing use and availability of free bioinformatics packages [16–20], normalization of functional genomic data is no longer a difficult task. It was likely deemed that the curation process is no longer of sufficient priority to the research community.

The predominant geographic origin of the GEO data has taken some interesting turns. Although the repository was becoming increasingly international, North American deposits once again began dominating after 2015. This was due to a sharp increase in North American deposits as well as a plateau in European ones. The reason behind these trends is not clear, but it is not likely the case that Europeans are now preferentially depositing on ArrayExpress, which continues to see only linear growth in their number of deposits [2].

Of note, it seems that North America is spearheading the sharp rise in HTS technologies in recent years, although its use is increasing in other regions as well. This finding is consistent with the fact that the U.S. has invested more in genomic research than any other country in the world [14]. HTS encompasses a variety of techniques, including ChIP-seq for genome binding profiling, and RNA-seq for transcriptome profiling. RNA-seq has some advantages over array-based technologies, being superior for detecting low-abundance transcripts, biologically distinct isoforms, and genetic variants [21, 22]. As sequencing becomes increasingly cheaper per base, and analysis software more widespread, RNA-seq may continue to overtake array-based technologies.

Interestingly, HTS was less likely to be used to study one of the six major disease areas. This suggests that HTS, as a relatively new technology, is still largely used for basic science and is still in the process of being adopted for more disease-specific applications (likely clinical studies). Nevertheless, this shifting technology carries important implications for the use of genomic data in clinical decisions and precision medicine [9, 23, 24]. Indeed, Array-based transcriptomics are already being used for cancer diagnosis [25, 26], staging and prognosis [27–30]. Moreover, the unique ability of RNA-seq to detect gene fusions and disease-associated isoforms appears to be an advantage for a clinical tool development [31], although comparatively few RNA-seq-based clinical tests currently exist [32, 33].

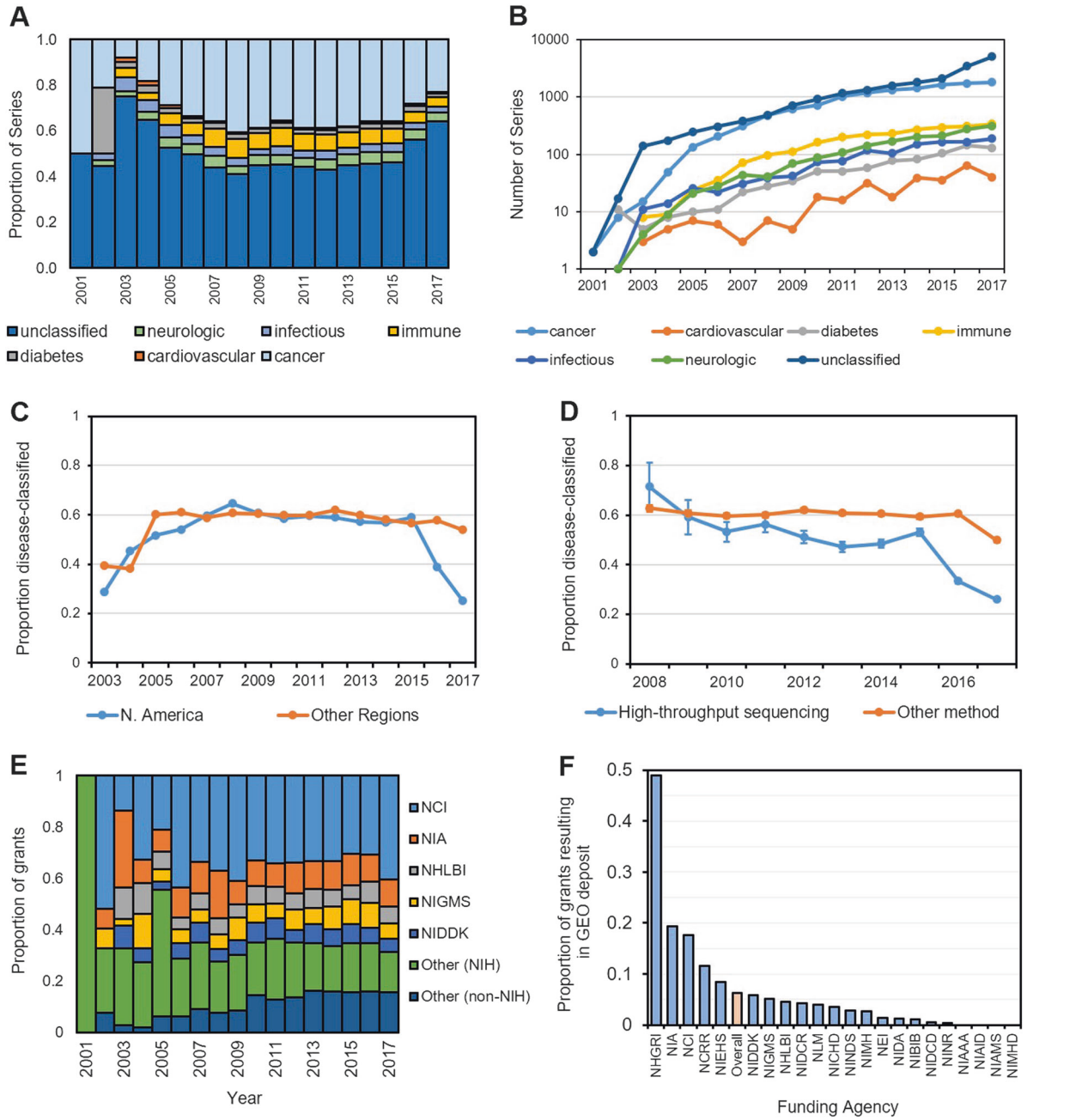

Fig. 5 Disease classifications and funding sources. **a** Proportion and **b** absolute number of Series studying one of six major disease areas over time. **c** Proportion of datasets classified to a major disease area, from North America vs. other regions. **d** Proportion of studies that classified to a major disease, for those using high-throughput sequencing vs. those that used other technologies. **e** Source of funding for human series with associated publications. **f** Ranked list of NIH institutions, by what proportion of grants produced a GEO deposit. Error bars represents standard error

As HTS becomes increasingly prevalent in the research world, clinicians will need to adapt so as to be able to effectively collect, analyze, and interpret data of such formats.

Of the investigated disease areas, there was a dominance of unclassified (likely basic research), cancer, and immunological diseases in the GEO deposits. The low percentages of GEO deposits in other disease areas, such as cardiovascular disease, may be concerning because the lack of sufficient human genomic data and understanding in the field may limit the development of genomics-based diagnostics and treatments [10, 31, 34, 35]. Related to this finding, the higher number of GEO Series per grant in select NIH institutes likely reflects a greater preference for and awareness of genomic data among the NIH-sponsored researchers. Perhaps a more interesting question is whether

Table 3 Association between human GEO deposits' geographic origin (North America vs. other regions) and their corresponding disease area (related to a major disease area or unclassified) from 2001 to 2017

	North America		Other regions				
Year	Related to major disease area	Unclassified	Related to major disease area	Unclassified	OR	95% CI	*P* value
2001	0	0	0	0	-	-	-
2002	19	16	2	1	0.59	0.05–7.17	>0.99
2003	44	109	13	20	0.62	0.28–1.36	0.297
2004	87	105	29	47	1.34	0.78–2.31	0.339
2005	168	157	86	57	0.71	0.48–1.06	0.107
2006	203	173	144	92	0.75	0.54–1.04	0.094
2007	312	210	199	139	1.04	0.79–1.37	0.831
2008	414	227	321	207	1.18	0.93–1.49	0.201
2009	522	338	447	292	1.01	0.83–1.23	0.959
2010	601	425	599	401	0.95	0.79–1.13	0.557
2011	814	552	768	514	0.99	0.84–1.15	0.874
2012	844	588	1026	630	0.88	0.76–1.02	0.09
2013	954	716	1103	741	0.90	0.78–1.02	0.107
2014	1001	757	1303	941	0.95	0.84–1.08	0.479
2015	1206	839	1419	1087	1.10	0.98–1.24	0.117
2016	1394	2192	1481	1080	0.46	0.42–0.51	6.84E−49
2017	1248	3709	1586	1353	0.29	0.26–0.32	3.46E−145
Total	9831 (25.2%)	11,113 (28.4%)	10,526 (26.9%)	7602 (19.5%)	0.64	0.61–0.67	5.02E−107

Note: Odds ratio (OR) represents the odds that a disease-specific GEO deposit originated from North America. The 6 major disease areas were cancer, cardiovascular diseases, diabetes, immunologic diseases, infectious diseases, and neurologic diseases, while all deposits not fitting in any of the six disease areas were categorized as unclassified. *CI* confidence intervals

Table 4 Association between the human GEO deposits' technology used (high-throughput sequencing vs. others) and their corresponding disease area (related to a major disease area or unclassified) from 2008 to 2017

	High-throughput sequencing		Other technology				
Year	Related to major disease area	Unclassified	Related to major disease area	Unclassified	OR	95% CI	*P* value
2008	15	6	720	428	1.49	0.57–3.86	0.499
2009	29	20	940	610	0.94	0.53–1.68	0.882
2010	81	71	1119	755	0.77	0.55–1.07	0.123
2011	149	116	1433	950	0.85	0.66–1.10	0.235
2012	211	202	1659	1016	0.64	0.52–0.79	2.99E-05
2013	280	313	1777	1144	0.58	0.48–0.69	1.39E-09
2014	466	497	1838	1201	0.61	0.53–0.71	5.32E-11
2015	644	571	1981	1355	0.77	0.68–0.88	1.27E-04
2016	1041	2079	1834	1193	0.33	0.29–0.36	1.31E-102
2017	1201	3425	1633	1636	0.35	0.32–0.39	6.76E-106
Total	4117 (11.2%)	7300 (19.9%)	14934 (40.8%)	10288 (28.1%)	0.39	0.37–0.41	1.00E-322

Note: Odds ratio (OR) represents the odds that a high-throughput sequencing study produced a disease-specific GEO deposit. The six major disease areas were cancer, cardiovascular diseases, diabetes, immunologic diseases, infectious diseases, and neurologic diseases, while all deposits not fit in any of the six disease areas were categorized as unclassified. No deposits using HTS were identified in the GEO prior to 2008. *HTS* high-throughput sequencing; *CI* confidence intervals

the research areas with fewer per-grant GEO deposits would need more genomic studies. This question may have profound clinical applications and present unique research opportunities. We found that cancer and basic science dominated GEO deposits, consistent with the largest funding sources (such as the NCI). On the other hand, endocrinological (diabetes, for example), neurological, and cardiovascular diseases lagged far behind. Due to

the faster accumulation of human genomic data and deeper understanding of cancer, cancer biologists, pathologists, and oncologists will more likely take advantage of genome-based diagnostics and targeted therapies than their colleagues in the fields with fewer genomic data deposits [4, 5, 36]. These advances in cancer will lead to more rapid and profound benefits for cancer patients.

In conclusion, we report increasing trends in GEO deposits (1) using HTS methods, (2) originating from North America, and (3) focusing on basic science applications. Cancer, immunological disease, and neurological diseases were the three disease areas with most deposits on the GEO. We also show that the NHGRI, NCI, NIEHS, and NCCR had a higher number of per-grant GEO Series than other NIH institutes and centers. More studies are needed to elucidate our observations. Our findings nonetheless may shed light on shaping future functional genomics-based research and clinical priorities.

Availability The database of human GSE, as well as code used for analysis, is available online at https://github.com/daniel-d-liu/GEO-Trends.

Funding The work was supported by an Initiative for Multidisciplinary Research Teams (IMRT) award from Rutgers University, Newark, NJ (to L.Z.).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991–995.
2. Kolesnikov N, Hastings E, Keays M, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res. 2015;43:D1113–1116.
3. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet. 2001;29:365–71.
4. Varmus H. The transformation of oncology. Science. 2016;352:123.
5. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375:1109–12.
6. Kahn SD. On the future of genomic data. Science. 2011;331:728–9.
7. Chin L, Hahn WC, Getz G, et al. Making sense of cancer genomic data. Genes Dev. 2011;25:534–55.
8. Varmus H. Genomic empowerment: the importance of public databases. Nat Genet. 2003;35(Suppl 1):3.
9. Zhang L. Biomarker discovery and validation in HCC diagnosis, prognosis, and therapy. In: Liu C, editor. Precision Molecular Pathology of Liver Cancer. Cham: Springer International Publishing, 2018. p. 95–113.
10. Lu M, Zhang J, Zhang L. Emerging concepts and methodologies in cancer biomarker discovery. Crit Rev Oncog. 2017;22(5-6):371-388. https://doi.org/10.1615/CritRevOncog.2017020626.
11. Kumar D. From evidence-based medicine to genomic medicine. Genom Med. 2007;1:95–104.
12. El-Sayed AM, Koenen KC, Galea S. Rethinking our public health genetics research paradigm. Am J Public Health. 2013;103(Suppl 1):S14–18.
13. Strauss KA, Puffenberger EG, Morton DH. One community's effort to control genetic disease. Am J Public Health. 2012;102:1300–6.
14. Pohlhaus JR, Cook-Deegan RM. Genomics research: world survey of public funding. BMC Genom. 2008;9:472.
15. Kim HJ, Fay MP, Feuer EJ, et al. Permutation tests for joinpoint regression with applications to cancer rates. Stat Med. 2000;19:335–51.
16. Zang C, Wang T, Deng K, et al. High-dimensional genomic data bias correction and data integration using MANCIE. Nat Commun. 2016;7:11305.
17. Li P, Piao Y, Shon HS, et al. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-seq data. BMC Bioinform. 2015;16:347.
18. Jiang Y, Oldridge DA, Diskin SJ, et al. CODEX: a normalization and copy number variation detection method for whole exome sequencing. Nucleic Acids Res. 2015;43:e39.
19. Chawade A, Alexandersson E, Levander F. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. J Proteome Res. 2014;13:3114–20.
20. Liang K, Keles S. Normalization of ChIP-seq data with control. BMC Bioinform. 2012;13:199.
21. Zhao S, Fung-Leung WP, Bittner A, et al. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9:e78644.
22. Kukurba KR, Montgomery SB. RNA sequencing and analysis. Cold Spring Harb Protoc. 2015;2015:951–69.
23. Basu A, Carlson JJ, Veenstra DL. A framework for prioritizing research investments in precision medicine. Med Decis Making. 2016;36:567–80.
24. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature. 2015;526:336–42.
25. Alexander EK, Kennedy GC, Baloch ZW, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. N Engl J Med. 2012;367:705–15.
26. Meiri E, Mueller WC, Rosenwald S, et al. A second-generation microRNA-based assay for diagnosing tumor tissue origin. Oncologist. 2012;17:801–12.
27. Mook S, Van't Veer LJ, Rutgers EJ, et al. Individualization of therapy using Mammaprint: from development to the MINDACT Trial. Cancer Genom Proteom. 2007;4:147–55.
28. Salazar R, Roepman P, Capella G, et al. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol. 2011;29:17–24.
29. Erho N, Crisan A, Vergara IA, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. PLoS One. 2013;8:e66855.
30. Knudsen BS, Kim HL, Erho N, et al. Application of a clinical whole-transcriptome assay for staging and prognosis of prostate cancer diagnosed in needle core biopsy specimens. J Mol Diagn. 2016;18:395–406.
31. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, et al. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet. 2016;17:257–71.
32. Sonu RJ, Jonas BA, Dwyre DM, et al. Optimal molecular methods in Detectingp190 (BCR-ABL) fusion variants in hematologic malignancies: a case report and review of the literature. Case Rep Hematol. 2015;2015:458052.

33. Doebele RC, Davis LE, Vaishnavi A, et al. An oncogenic NTRK fusion in a patient with soft-tissue sarcoma with response to the tropomyosin-related kinase inhibitor LOXO-101. Cancer Discov. 2015;5:1049–57.
34. Bertagnolli MM, Sartor O, Chabner BA, et al. Advantages of a truly open-access data-sharing model. N Engl J Med. 2017;376: 1178–81.
35. Van Voorhis WC, Adams JH, Adelfio R, et al. Open source drug discovery with the malaria box compound collection for neglected diseases and beyond. PLoS Pathog. 2016;12: e1005763.
36. Scott JG, Berglund A, Schell MJ, et al. A genome-based model for adjusting radiotherapy dose (GARD): a retrospective, cohort-based study. Lancet Oncol. 2017;18:202–11.