Network Analysis for Directing Standardization Efforts in Genomics Using MEDLINE

Eun Bit Bae
Institute of Quantum Biophysics, Sungkyunkwan University
Se Jin Nam
Sungkyunkwan University
Sungin Lee
Institute of Quantum Biophysics, Sungkyunkwan University
Sun-Ju Ahn (ahnsunju@skku.edu)
Institute of Quantum Biophysics, Sungkyunkwan University

Research Article

Keywords: Network analysis, Standard, Genomics, Genomic sequence, Keyword analysis, PubMed

Posted Date: February 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1245232/v1

Abstract

Background: Biotechnology in genomics, such as sequencing devices and gene quantification software, has proliferated and been applied to clinical settings. However, the lack of standards applicable to it poses practical problems in interoperability and reusability of the technology across various application domains. This study aims to visualize and identify the standard trends in clinical genomics and to suggest areas on which standardization efforts must focus.

Methods: Of 16,538 articles retrieved from PubMed, published from 1975 to 2020, using search keywords “genomics and standard” and “clinical genomic sequence and standard”, terms were extracted from the abstracts and titles of 15,855 articles. Our analysis includes 1) keyword-based network analysis, considering the characteristics of keyword clusters; 2) period analysis with six phases; 3) statistical analysis; 4) content analysis.

Results: Our research trend showed an increasing trend from 2003, years marked by the completion of the human genome project (2003). The content analysis showed that keywords related to such concepts as gene types for analysis, and analysis techniques were increased in phase 3 when US-FDA first approved the next-generation sequencer. During 2017-2019, oncology-relevant terms were clustered and contributed to the increasing trend in phase 4 of the content analysis. In the statistical analysis, all the categories showed high regression values (R^2>.586) throughout the whole analysis period and phase-based statistical analysis showed significance only in Genetics terminology category (P=.039*) at phase 4.

Conclusions: This study investigated, through the network and statistical analysis, trend shifts in relation to clinical genomics. The shifts offer a solid basis for standards development that shows where future standardization opportunities and gaps lie.

Introduction

The dawn of the 20th century saw the rise of medical genetics research on humans due to the discovery of Mendelian inheritance disorders [1, 2]. Remarkable progress in medical genetics has been made in the latter part of the 20th century, notably in cancer genetics [3]. Research on disease diagnosis using genomic sequencing technologies has gained momentum, thanks to the wide availability of next-generation sequencing (NGS) methods.

Guidelines and standards - industry-developed or originating from international standards development bodies – are for clinical examination and diagnosis of diseases such as cancer, leukemia, and tuberculosis [4–7]. These standards, from such organizations as the American College of Medical Genetics (ACMG), Association of Molecular Pathology, Microarray Quality Control Consortium [4, 7–9], have enabled a variety of sequencing technologies and methods [7]

However, extant standards and their coverage could not be claimed to be sufficient to meet the standardization demand from the market, notably evident in the clinical applications of NGS to disease diagnosis [4–7]. For the genomics market to thrive, it is of significant import for international standards communities to be able to scan the genomics environment, gather necessary technical and human resources for standardization, to distill and refine priorities for genomics standardization. One way to help scan the genomics environment is to apply network analysis on the artifacts of research articles to reveal environmental changes that can be used as guidance for standardization efforts.

To explore specific research trends, network analysis using bibliometric data has been widely used and applied to various research domains, for example, genomics [10], public health [11, 12] and medicine [13]. Network analysis in this study is used to divulge research trend changes. The identification of such trend changes can enable standards development communities to formulate a strategic roadmap to meet the standardization demand from the genomics market. In detail, this study uses network analysis 1) to suggest background knowledge on recent genomics’ trends and topics in standardization, 2) to examine temporal trends and related critical development which drives changes in trend, that is, any development that acts as major factors and indicators to which standards development communities should pay attention.

Methods

Study Flow

The overall study procedure is shown in [Supplementary file, Figure S1] and summarized as follows: 1) search articles with two Medical Subject Heading (MeSH) terms (“genomics and standard” and “clinical genomic sequence and standard”) in PubMed; 2) export PMID numbers; 3) extract keywords from the abstracts and titles of the articles; 4) keyword refinement; 5) development of a similarity table between keywords; 6) development of a similarity table between similarities based on publication years; and 7) classification of keywords into two groups of categories for statistical analysis.

Data Source

We used articles in Pubmed indexed with two MeSH terms, “genomics and standard” and “clinical genomic sequence and standard”, published between 1975 and September 2020. The search returned 16,550 articles that contained various types of research papers, such as reviews, original articles, and perspectives. Of the articles, 10,000 articles were indexed with the search term “Genomics and standard”, and 6,550 articles with "Clinical genomic sequence and standard”. Of the 16,550 articles, we used 15,855 articles whose abstracts and titles were accessible and written in English.

Data Preparation

A total of 84,644 keywords were extracted from 15,855 articles. The keywords were extracted using the TextRank algorithm [14] from the abstracts and titles of the articles.

To refine the keywords, a set of exclusion criteria was established, based on the consensus of two subject experts. The criteria are 1) non-technical terms with little or no relevance to the study such as “happy” and “very”; 2) temporal terms such as months, weekdays, as well as other temporal terms that, used alone, do not provide precise temporal context, such as “April” (instead of “April 2004”) or “Monday”; and 3) compound nouns with two conditions: a) frequencies of a compound noun of whose constituent terms have been already counted individually AND b) the compound noun does not constitute a meaningful term such as “genomics proteomics” and “protein gene”.

Terms merging criteria was as follows: 1) abbreviation – a spelled-out word and its abbreviation are merged to the latter; 2) plural and singular nouns – when both exist for a word, merge them to the singular. Terms that appeared less than 12 times were deleted. Most of the deleted terms comprised compound nouns with more than 3 words, which rendered them meaningless.

The keyword refinement resulted in a list of 330 keywords, with the combined frequency of N = 16,213. The data preparation described above is shown in [Supplementary file, Figure S1].

Network Analysis

In network analysis, a network consists of nodes and edges. And a node represents a keyword and an edge relatedness between two keywords. In this study, the relatedness between two keywords is calculated by weighted Jaccard similarity using frequencies of the keywords [15, 16]. The weight (that is, importance) of a node in the network was determined by the PageRank algorithm. And a community detection algorithm was used to cluster keywords.

Similarity between Keywords

To calculate the similarity between two keywords, a weighted Jaccard algorithm was used.

First, a two-dimensional annual frequency matrix was generated with a frequency of each term by publication years - a matrix of 330(the number of keywords) x 46 (the number of publication years, from 1975 to 2020). In the following equation, S and T

represent two keywords. The relationship of the keywords in an article is calculated as

$$J(S, T) = \frac{\sum_K \min(S_\square, T_\square)}{\sum_K \max(S_\square, T_\square)}$$

In this study, Gephi 0.8.2. was used to create a network model of the relationships between keywords.

Period Analysis

To identify inflection points of keyword similarity, we conducted three rounds of similarity analysis (Fig. 2). We submit that an infection in similarity provides an indicator that there has been a significant development or event that deserves the attention of standards development communities.

1) Similarity-based on frequency values of two publication years (for example, years 2000 and 2001)

2) Similarity-based on similarity values between two two-year blocks (for example, years 2000–2001 and 2001–2002)

3) Similarity between similarity values between two three-year blocks (for example, years 2000–2002 and 2001–2003).

Please note that phase 0 (1975–1999) was not included in the analysis, due to the low-frequency values (frequency of 10 to 72).

Content Analysis

The keywords were classified into two sets of categories, CAT1 and CAT2: 1) CAT1 contains six categories: Biology, General, Genetics, Medicine, Proteomics, and Statistics; and 2) CAT2 contains 12 categories: Biologicals/Metabolics, Clinical, Company/ Consortium, Database/ Software, Disease, Gene, Genetics term, Methods, Organism, Pathogen, Proteomics, and Statistics.

Statistical Analysis

To examine the relationship between phases, linear regression analysis was performed with keyword frequencies and publication years for each category of CAT1 and CAT2. To statistically estimate research trends, we constructed a univariate generalized linear model within phases for each category of CAT1 and CAT2. SPSS Statistics ver.26, IBM was used for the statistical analysis.

Results

The network analysis

A total of eight clusters were created as shown in Fig. 1, and a sample of PageRank scores of each cluster is listed in Table 1.

Table 1
Keywords in clusters.

Cluster 0 (C0)		C1		C2		C3	
Keyword	PageRank	Keyword	PageRank	Keyword	PageRank	Keyword	PageRank
Genome	0.0167	gene	0.0244	nano gram	0.0159	WGS	0.0087
SNP	0.0148	mRNA	0.0067	tumor	0.0105	Escherichia	0.0084
Disease	0.0128	qPCR	0.0066	therapy	0.0076	bacteria	0.0061
Allele	0.0119	microarray	0.0052	EGFR	0.0056	pathogen	0.0053
Clinician	0.0097	Arabidopsis	0.0048	IHC	0.0043	Mycobacterium	0.0051
Genomics	0.0084	geNorm	0.0048	KRAS	0.0041	MLST	0.0047
Illumina	0.0075	gene normalization	0.0047	NSCLC	0.0041	NCBI	0.0039
Bayesian	0.0072	NormFinder	0.0045	targeted therapy	0.0040	MiSeq	0.0038
genetics	0.0058	cDNA	0.0042	amplicon	0.0037	Pseudomonas	0.0038
bioinformatics	0.0058	miRNA	0.0035	tumor DNA	0.0037	Streptococcus	0.0037
C4		C5		C6		C7	
Keyword	PageRank	Keyword	PageRank	Keyword	PageRank	Keyword	PageRank
CpG	0.0053	rRNA	0.0042	protein	0.0152	diagnosis	0.0077
WHO	0.0051	nucleotide	0.0038	biomarker	0.0095	CNV	0.0056
DNA methylation	0.0040	GenBank	0.0037	proteomics	0.0087	genomic hybridization	0.0054
methylation	0.0037	codon	0.0036	algorithm	0.0075	genomic DNA	0.0047
MGMT	0.0036	genotyping	0.0033	peptide	0.0063	STR	0.0032
inhibitor	0.0036	mitochondrial genome	0.0031	database	0.0060	chromosome	0.0031
AML	0.0036	mtDNA	0.0031	knowledge	0.0047	BAC	0.0030
ROC	0.0034	phylogenetic	0.0031	reproducibility	0.0046	aCGH	0.0029
TMZ	0.0032	tRNA	0.0031	FDR	0.0040	MLPA	0.0028
IDH	0.0031	RNA	0.0029	measurement	0.0040	haplotype	0.0028

In C0, terms related to genetic materials (i.e., “genome”, “SNP”), clinical terminology (i.e., “disease”), and technology (i.e., “Illumina”, “Bayesian”) are clustered. In C1, genetic materials (“gene”, “mRNA”, and “cDNA”.) and gene analysis techniques (“qPCR”, and “Normfinder”) are clustered. In C2, the term “nano gram” and oncology-related keywords (“tumor”, “IHC”, “NSCLC” and “tumor DNA”) are grouped together. In C3, terms related to pathogens are clustered (“Escherichia”, “bacteria”, “pathogen”, and “Mycobacterium”). In C4, DNA methylation-related terms (“CpG”, “DNA methylation”, “methylation”, “MGMT”) and in C5, gene-related or phylogenetic terms (“mitochondrial genome”, “mtDNA”, “phylogenetic”, “tRNA”, and “RNA”) were clustered. In C6, proteomics terms are clustered (“protein”, “biomarker”, “proteomics”, “algorithm”, and “peptide”). The last cluster, C7, genetic technology and clinic-relevant terms are present (“diagnosis”, “CNV”, “genomic hybridization”).

Period Analysis based on publication years

We identified three main inflection points for each similarity analysis. The inflection points signify changing trends: in Fig. 2A, the infection points emerged in 2003:2004 (Similarity = .294), 2012:2013 (S = .485), and 2017:2018 (S = .541) where the trend has started to plateau (Table 2). And in Fig. 2B, inflection points in 2002–2003:2003–2004 (S = .518); 2011–2012:2012–2013 (S = .684); and 2016–2017:2017–2018 (S = .736) were identified, while, in Fig. 2C, 2001–2003:2002–2004 (S = .612); 2010–2012:2011–2013 (S = .770); and 2015–2017:2016–2018 (S = .798) were observed. The similarity scores for each period analysis are shown in Table 2.

Table 2
Similarity results based on different year ranges

Similarity (1 year)		**Similarity (2 years)**		**Similarity (3 years)**	
Year	Similarity	Year	Similarity	Year	Similarity
2000–2001	0.267	2000:2001–2001:2002	0.515	2000:2002 – 2001:2003	0.649
2001–2002	0.274	2001:2002–2002:2003	0.572	2001:2003 – 2002:2004	0.612
2002–2003	0.298	2002:2003–2003:2004	0.518	2002:2004 – 2003:2005	0.652
2003–2004	0.294	2003:2004–2004:2005	0.566	2003:2005 – 2004:2006	0.695
2004–2005	0.396	2004:2005–2005:2006	0.622	2004:2006 – 2005:2007	0.735
2005–2006	0.377	2005:2006–2006:2007	0.624	2005:2007 – 2006:2008	0.710
2006–2007	0.389	2006:2007–2007:2008	0.605	2006:2008 – 2007:2009	0.717
2007–2008	0.387	2007:2008–2008:2009	0.628	2007:2009 – 2008:2010	0.732
2008–2009	0.393	2008–2009–2009:2010	0.651	2008:2010 – 2009:2011	0.745
2009–2010	0.434	2009:2010–2010:2011	0.640	2009:2011 – 2010:2012	0.708
2010–2011	0.439	2010:2011–2011:2012	0.624	2010:2012 – 2011:2013	0.770
2011–2012	0.397	2011:2012–2012:2013	0.684	2011:2013 – 2012:2014	0.761
2012–2013	0.485	2012:2013–2013:2014	0.692	2012:2014 – 2013:2015	0.757
2013–2014	0.486	2013:2014–2014:2015	0.682	2013:2015 – 2014:2016	0.766
2014–2015	0.480	2014:2015–2015:2016	0.689	2014:2016 – 2015:2017	0.775
2015–2016	0.491	2015:2016–2016:2017	0.716	2015:2017 – 2016:2018	0.798
2016–2017	0.503	2016:2017–2017:2018	0.736	2016:2018 – 2017:2019	0.796
2017–2018	0.541	2017:2018–2018:2019	0.737	2017:2019 – 2018:2020	0.774
2018–2019	0.544	2018:2019–2019:2020	0.675		
2019–2020	0.349				

Content Analysis

The combined frequencies of keywords belonging to each category of CAT1 and CAT 2 are computed. Each keyword belongs to only one category.

Genetics in CAT 1 has the highest frequency (n = 8,777, 54.1%) in CAT1, followed by. Medicine (n = 2,856, 17.6%), Proteomics (n = 2,257, 13.9%), General (n = 992, 6.1%), Biology (n = 707, 4.3%), and Statistics (n = 624, 3.81%).

Gene in CAT2 has the highest frequency (n = 3276, 20.2%), followed by Genetics terminology (n = 3019, 18.6%), Methods (n = 1725, 10.6%), Database/Software (n = 1393, 8.59%), Disease (n = 1204, 7.42%), Clinical (n = 1103, 6.8%), Proteomics (n = 1034, 6.37%), Pathogen (n = 1006, 6.2%), Statistics (n = 720, 4.44%), Biologicals (n = 707, 4.36%), Company/Consortium (n = 536, 3.3%), and Organism (n = 490, 3.02%).

We examined the trend of each term from phase 0 to phase 4 in CAT2 as follows:

In [Supplementary file, Figure S2], “Escherichia” showed the highest frequency in phase 2, and “Mycobacterium” in phase 4. In Statistics, “Bayesian” and “algorithm” were of the highest frequency in phase 2, while the frequency of the latter steadily decreased until phase 4. The frequency of “Bayesian” increased from phase 3 to 4.

In the Company/Consortium graph, “Illumina,” “Taqman” were of the highest frequency at phase 4, and “Illumina” and “ACMG” showed an increasing trend during the whole period. In Database, the term “bioinformatics” showed the highest frequency at phase 4. In Gene, the terms “gene”, “genome”, “allele”, “codon”, “cDNA”, “chromosome”, “DNA”, and “mtDNA” exhibited the highest frequencies at phase 2 and started to decrease in frequency from phase 3 to phase 4.

Terms denoting relatively smaller gene fragments, such as “RNA”, “miRNA”, “rRNA”, “exome”, “tRNA”, showed an increasing trend from phase 3 to 4. In Software, terms referring to gene quantification software, “NormFinder”, “geNorm”, and “BestKeeper”, were highest in frequency at phase 3 and “ClinGen” showed an increasing trend from phase 3 to 4. In Methods, “WGS”, “GWAS”, and “MiSeq” exhibited an increasing trend from phase 2 and peaked in frequency at phase 4.

On the other hand, “microarray,” “genomic hybridization,” and “gene microarray” showed the highest frequency in phase 2, and “qPCR” peaked in frequency in phase 3. In Clinical, “Clinician”, “therapy”, “diagnosis”, “precision”, “targeted therapy”, and “biopsy” all showed an increasing trend until phase 4; and in Disease, the term “disease” and oncology-related terms, such as “tumor”, “NSCLC”, “AML”, “GBM”, “tumor DNA”, and “adenocarcinoma” showed an increasing trend throughout the phases.

Statistical Analysis

Linear Regression without Phase

To evaluate linear trends, linear regression was conducted with keyword frequencies for publication years from 1975 to 2020. Although 2020 showed a decreasing trend in CAT1 and CAT2, all the categories in CAT1 and CAT 2 showed high regression values (from 0.586 (Company/Consortium) to 0.764 (Biology)) as shown in Table 3, Fig. 3. All the categories showed an increasing linear correlation between keyword frequencies and publication years.

Table 3
Linear regression based on keyword frequency in CAT1 and CAT2

	Category	R^2
CAT1	Biology	0.764
	General	0.587
	Genetics	0.717
	Medicine	0.653
	Proteomics	0.673
	Statistics	0.666
CAT2	Clinical	0.657
	Company/Consortium	0.586
	Database/software	0.684
	Disease	0.625
	Gene	0.664
	Genetics terminology	0.740
	Metabolite/Biologicals	0.652
	Methods	0.736
	Organism	0.741
	Pathogen	0.737
	Proteomics	0.678
	Statistics	0.648

Generalized Linear Model within Phase

The linear regression analysis without phase demonstrated a high correlation ($R^2 > .586$). To analyze phase-based linear analysis for each category, we performed General Linear Model (GLM) evaluation based on phases (Fig. 4, Table 4). There was no linear correlation found in CAT1 categories (Table S1) while correlations were observed in several CAT2 categories (Table 4): Gene ($P = .003$) and Pathogen ($P = .030$) were statistically significant in phase 0, and Gene ($P = .004$) and Proteomics ($P = .044$) were statistically significant in phase 1. In phase 2, only Proteomics ($P = .001$) was significant. In phase 3, Proteomics ($P = .045$) and Software ($P = .004$) were significant, and in phase 4, only Genetics terminology was significantly fitted with the linear model ($P = .039$).

Table 4
Generalized linear model results of CAT2 from phase 0 to phase 4.

Phase	Category	B	SE	t	Sig.	95% Confidence Interval	
						Lower	Upper
Phase 0	Biologicals	0.647	0.911	0.710	0.478	-1.146	2.439
	Clinical	0.004	0.875	0.005	0.996	-1.717	1.726
	Company/Institute	0.022	1.078	0.020	0.984	-2.100	2.143
	Data related	0.089	0.991	0.090	0.928	-1.861	2.040
	Disease	0.739	0.868	0.851	0.395	-0.969	2.446
	Gene	2.347	0.781	3.007	0.003**	0.812	3.883
	Genetics terminology	0.299	0.783	0.382	0.703	-1.242	1.841
	Methods	0.504	0.763	0.661	0.509	-0.996	2.005
	Organism	2.338	1.226	1.907	0.057	-0.074	4.749
	Pathogen	2.036	0.933	2.182	0.030*	0.200	3.873
	Proteomics	1.160	1.180	0.983	0.326	-1.161	3.481
	Software	-0.315	1.281	-0.246	0.806	-2.835	2.205
Phase 1	Biologicals	0.146	1.305	0.112	0.911	-2.422	2.714
	Clinical	0.876	1.254	0.698	0.485	-1.591	3.342
	Company/Institute	0.935	1.544	0.606	0.545	-2.103	3.974
	Data related	0.849	1.420	0.598	0.550	-1.944	3.643
	Disease	1.320	1.243	1.062	0.289	-1.125	3.765
	Gene	3.256	1.118	2.912	0.004**	1.056	5.456
	Genetics terminology	0.798	1.122	0.711	0.477	-1.410	3.006
	Methods	0.487	1.093	0.445	0.656	-1.663	2.636
	Organism	2.653	1.756	1.511	0.132	-0.801	6.108
	Pathogen	2.320	1.337	1.735	0.084	-0.311	4.951
	Proteomics	3.420	1.690	2.024	0.044*	0.095	6.745
	Software	-0.555	1.835	-0.303	0.762	-4.165	3.055
Phase 2	Biologicals	-0.318	8.863	-0.036	0.971	-17.757	17.120
	Clinical	-1.951	8.514	-0.229	0.819	-18.704	14.801
	Company/Institute	3.622	10.490	0.345	0.730	-17.017	24.260
	Data related	3.631	9.644	0.376	0.707	-15.343	22.605
	Disease	1.089	8.441	0.129	0.897	-15.519	17.697
	Gene	13.437	7.594	1.769	0.078	-1.504	28.377
	Genetics terminology	8.486	7.622	1.113	0.266	-6.511	23.483

	Methods	1.530	7.421	0.206	0.837	-13.070	16.131
	Organism	8.160	11.925	0.684	0.494	-15.303	31.623
	Pathogen	3.541	9.080	0.390	0.697	-14.325	21.407
	Proteomics	38.460	11.478	3.351	0.001**	15.877	61.043
	Software	11.910	12.461	0.956	0.340	-12.607	36.427
Phase 3	Biologicals	0.421	7.619	0.055	0.956	-14.570	15.412
	Clinical	1.493	7.319	0.204	0.838	-12.908	15.894
	Company/Institute	3.468	9.017	0.385	0.701	-14.274	21.209
	Data related	1.631	8.290	0.197	0.844	-14.680	17.941
	Disease	3.874	7.256	0.534	0.594	-10.403	18.151
	Gene	11.117	6.528	1.703	0.090	-1.726	23.961
	Genetics terminology	11.921	6.552	1.819	0.070	-0.971	24.813
	Methods	-0.081	6.379	-0.013	0.990	-12.632	12.471
	Organism	4.938	10.251	0.482	0.630	-15.232	25.107
	Pathogen	4.255	7.806	0.545	0.586	-11.103	19.614
	Proteomics	19.860	9.867	2.013	0.045*	0.446	39.274
	Software	30.785	10.712	2.874	0.004**	9.709	51.861
Phase 4	Biologicals	2.348	8.725	0.269	0.788	-14.818	19.514
	Clinical	9.000	8.381	1.074	0.284	-7.490	25.490
	Company/Institute	4.385	10.326	0.425	0.671	-15.931	24.700
	Data related	6.294	9.493	0.663	0.508	-12.383	24.971
	Disease	7.179	8.309	0.864	0.388	-9.170	23.527
	Gene	10.745	7.475	1.437	0.152	-3.963	25.452
	Genetics term	15.587	7.503	2.077	0.039*	0.824	30.350
	Methods	1.537	7.305	0.210	0.833	-12.835	15.909
	Organism	5.778	11.738	0.492	0.623	-17.318	28.873
	Pathogen	6.952	8.938	0.778	0.437	-10.634	24.539
	Proteomics	11.700	11.299	1.036	0.301	-10.530	33.930
	Software	15.750	12.256	1.284	0.200	-8.384	39.884

Discussion

In this study, we have investigated the trends in clinical genetics from 1975 to 2020.

We analyzed term characteristics by each cluster, such as oncology-related cluster (C2). The period and content analysis revealed notable trend changes in clinical genomics, especially in gene analysis technology and target genes to be analyzed in 2003 and 2013.

What has contributed to these changes is of significance for the standards development communities, since standardization efforts should conflate these changes in gathering standard resources, planning, and developing future standards for genomics to meet the ever-changing demands from the domain and stay ahead.

For instance, in April 2003, the Human Genome Project, the world's largest collaborative biological project from 1990, was completed [17], ramifications of which seemed to have been reflected in the trend shift at phase 2, and which, further down the road, contributed to the increased occurrences of genetic analysis terms such as "qPCR", "microarray", "electrophoresis", and "Taqman".

Another example may be gleaned from an event in 2013, the approval of Illumina's sequencer by US-FDA [18] in 2013. An increasing trend shift was observed at phase 3 in the form of increased frequencies of sequencing-related terms ("miRNA", "rRNA"), devices ("Illumina", "MiSeq"), and methods ("WGS", "GWAS", "geNorm", "NormFinder").

Other shifts of note are: 1) In phase 4, Genetics terminology category showed a significant increasing linear trend. From content analysis, we identified drastic increasing trends in clinical terminology, such as "clinician", "therapy", "diagnosis", "precision", and "pathogen", especially oncology-related terminology, such as "tumor", "NSCLC", "AML", "GBM", and "tumor DNA". 2) From phase 1 to phase 3, there was a trend shift in terms related to gene analysis technology and target genes, with increased occurrences of terms for smaller size genes from the large ones (i.e., from "genomic DNA", "DNA" and "chromosomes" to "RNA", "miRNA", "rRNA", "exome", "tRNA").

Clinical genomics is a living subject, subject to a rethinking of truth based on new discoveries. And standard development communities' endeavor to stay ahead of the most current research and development in the field. Unfortunately, the development of a standard tends to lag the current development of the domain which the standard aims to address. The standard development communities need guidance that helps map out the domain for environmental scanning, distillation, and refinement of urgently required standards for the domain to flourish. The trend shifts of this study provide such guidance.

Limitation

This study has limitations in two areas: 1) although two experts agreed to classify studied keywords into two sets of categories with iterative refinement, it was practically out of reach to achieve definitive and precise categorization; and 2) articles past September 2020 were not included in our study, which may or may not has influenced the downward trend in 2020.

Conclusions

The effective clinical application of genomics research and development demands standardization, without which significant achievements may not be used as widely as they can be. This study provides analysis results that can be used as guidance for directing future standards development efforts in clinical genomics.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files

Conflicts of interest

The authors declare that they have no competing interests.

Funding

This project was supported by the Korean Ministry of Trade, Industry, and Energy (No. 20011748, No. 20012610).

Authors' contributions

Sun-Ju Ahn initiated the study. Sun-Ju Ahn, Eun Bit Bae & Se Jin Nam designed methodology; Eun Bit Bae & Se Jin Nam constructed overall study design, detailed study outline and prepared keyword data; Se Jin Nam conducted network analysis; Eun Bit Bae conducted the statistical analysis, interpreted results and led the writing of the manuscript. Eun Bit Bae & Sungin Lee reviewed and finalized the manuscript.

Acknowledgements

Not applicable

References

1. C. W.E., Mendel's Law of Heredity, Proceedings of the American Academy of Arts and Sciences, 39 (1903) 223-242.
2. E.D. Wills, Biochemical genetics: inborn errors of metabolism, Biochemical Basis of Medicine, Elsevier1985, pp. 511-520.
3. S. Hodgson, Advances in cancer genetics, Clin Med, 9 (2009) 151-153.
4. E.M. McCormick, M.T. Lott, M.C. Dulik, L. Shen, M. Attimonelli, O. Vitale, A. Karaa, R. Bai, D.E. Pineda-Alvarez, L.N. Singh, C.M. Stanley, S. Wong, A. Bhardwaj, D. Merkurjev, R. Mao, N. Sondheimer, S. Zhang, V. Procaccio, D.C. Wallace, X. Gai, M.J. Falk, Specifications of the ACMG/AMP standards and guidelines for mitochondrial DNA variant interpretation, Hum Mutat, 41 (2020) 2028-2057.
5. B.D. Cheson, J.M. Bennet, K.J. Kopecky, T. Büchner, C.L. Willman, E.H. Estey, C.A. Schiffer, H. Doehner, M.S. Tallman, T.A. Lister, F. Lo-Coco, R. Willemze, A. Biondi, W. Hiddemann, R.A. Larson, B. Löwenberg, M.A. Sanz, D.R. Head, R. Ohno, C.D. Bloomfield, Revised recommendations of the International Working Group for Diagnosis, Standardization of Response Criteria, Treatment Outcomes, and Reporting Standards for Therapeutic Trials in Acute Myeloid Leukemia, Journal of clinical oncology, 21 (2003) 4642-4649.
6. C.J. Meehan, G.A. Goig, T.A. Kohl, L. Verboven, A. Dippenaar, M. Ezewudo, M.R. Farhat, J.L. Guthrie, K. Laukens, P. Miotto, B. Ofori-Anyinam, V. Dreyer, P. Supply, A. Suresh, C. Utpatel, D. van Soolingen, Y. Zhou, P.M. Ashton, D. Brites, A.M. Cabibbe, A. Van Rie, Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. , 17(9), 533–545. https://doi.org/10.1038/s41579-019-0214-5, Nature reviews. Microbiology, 17 (2019) 533-545.
7. C.E. Mason, E. Afshinnekoo, S. Tighe, S. Wu, S. Levy, International Standards for Genomes, Transcriptomes, and Metagenomes, J Biomol Tech, 28 (2017) 8-18.
8. <WHO-CDS-TB-2018.19-eng.pdf>.
9. G. Matthijs, E. Souche, M. Alders, A. Corveleyn, S. Eck, I. Feenstra, V. Race, E. Sistermans, M. Sturm, M. Weiss, H. Yntema, E. Bakker, H. Scheffer, P. Bauer, EuroGentest, G. European Society of Human, Guidelines for diagnostic next-generation sequencing, Eur J Hum Genet, 24 (2016) 2-5.

10. M. Ernst, Y. Du, G. Warsow, M. Hamed, N. Endlich, K. Endlich, H. Murua Escobar, L.M. Sklarz, S. Sender, C. Junghanss, S. Moller, G. Fuellen, S. Struckmann, FocusHeuristics - expression-data-driven network optimization and disease gene prediction, Sci Rep, 7 (2017) 42638.
11. D. Gu, T. Li, X. Wang, X. Yang, Z. Yu, Visualizing the intellectual structure and evolution of electronic health and telemedicine research, Int J Med Inform, 130 (2019) 103947.
12. C. Liang, S. Qiao, B. Olatosi, T. Lyu, X. Li, Emergence and evolution of big data science in HIV research: Bibliometric analysis of federally sponsored studies 2000-2019, Int J Med Inform, 154 (2021) 104558.
13. R. Takahashi, Y. Kajikawa, Computer-aided diagnosis: A survey with bibliometric analysis, Int J Med Inform, 101 (2017) 58-67.
14. R.T. Mihalcea, Paul, TextRank: Bringing Order into Text, Empirical Methods in Natural Language Processing, Association for Computational Linguistics, EMNLP | WS, 2004, pp. 404-411.
15. Y.-J. Son, Lee, S-K., Nam, SJ., Shim, J.L., Exploring Research Topics and Trends in Nursing-related Communication in Intensive Care Units Using Social Network Analysis, CIN, 36 (2018) 383-392.
16. Y.A. Ji, S.J. Nam, H.G. Kim, J. Lee, S.K. Lee, Research topics and trends in medical education by social network analysis, BMC Med Educ, 18 (2018) 222.
17. N. Riley, Out of date: genetics, history and the British novel of the 1990s, Med Humanit, (2021).
18. F.S. Collins, Hamburg, Margaret A., First FDA Authorization for Next-Generation Sequencer, N Engl J Med, 369 (2013) 2369-2371.

Figures

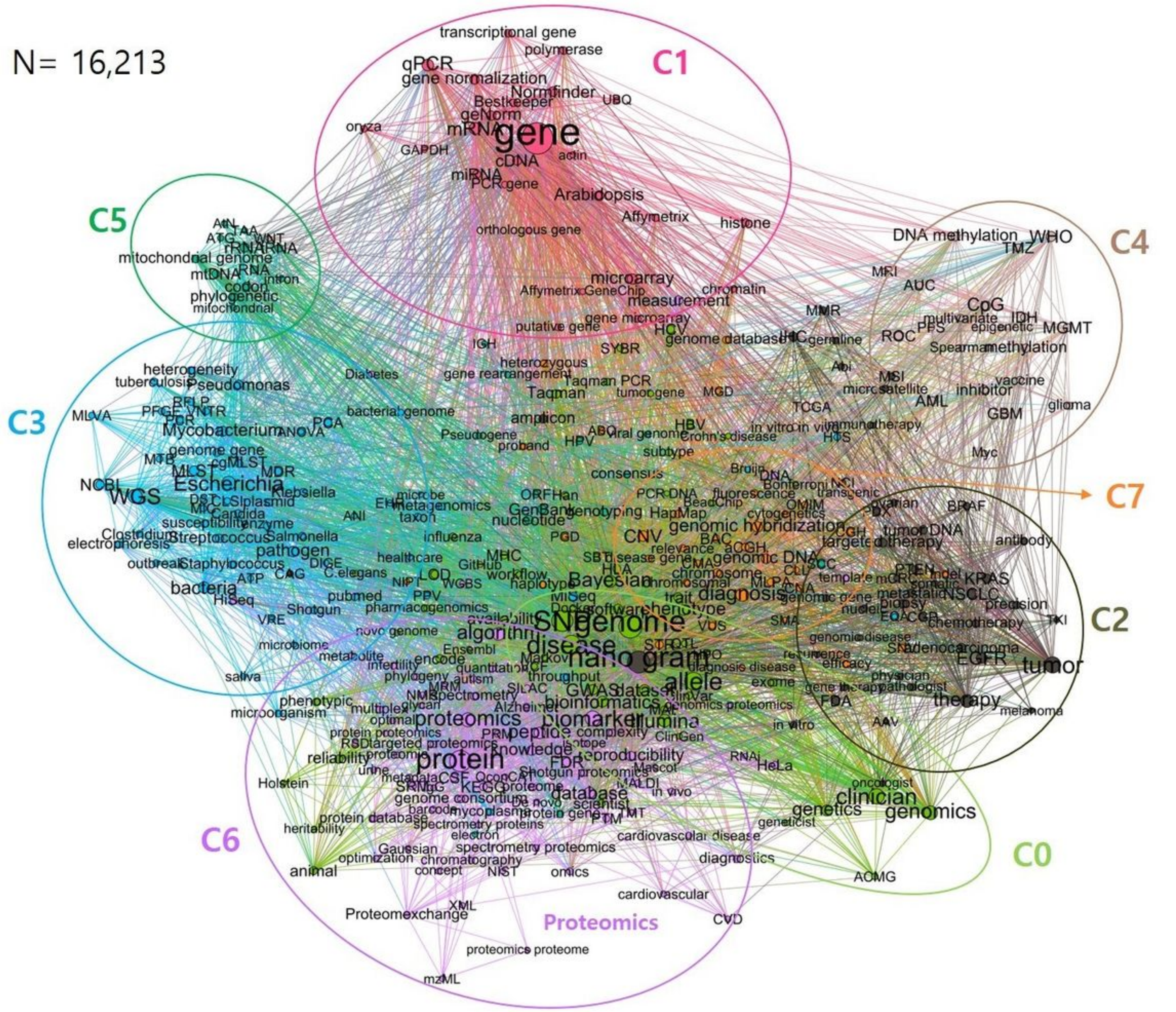

Figure 1

Network connectivity between keywords for the total period (1975–2020). In the network, the total frequency of 338 keywords is 16,213. The color of an edge represents the same similarity value and represents the cluster. Each keyword represents a node, and may have many edges to and from other keywords. A node size represents a PageRank score. Each color group represents a distinct cluster of keywords.

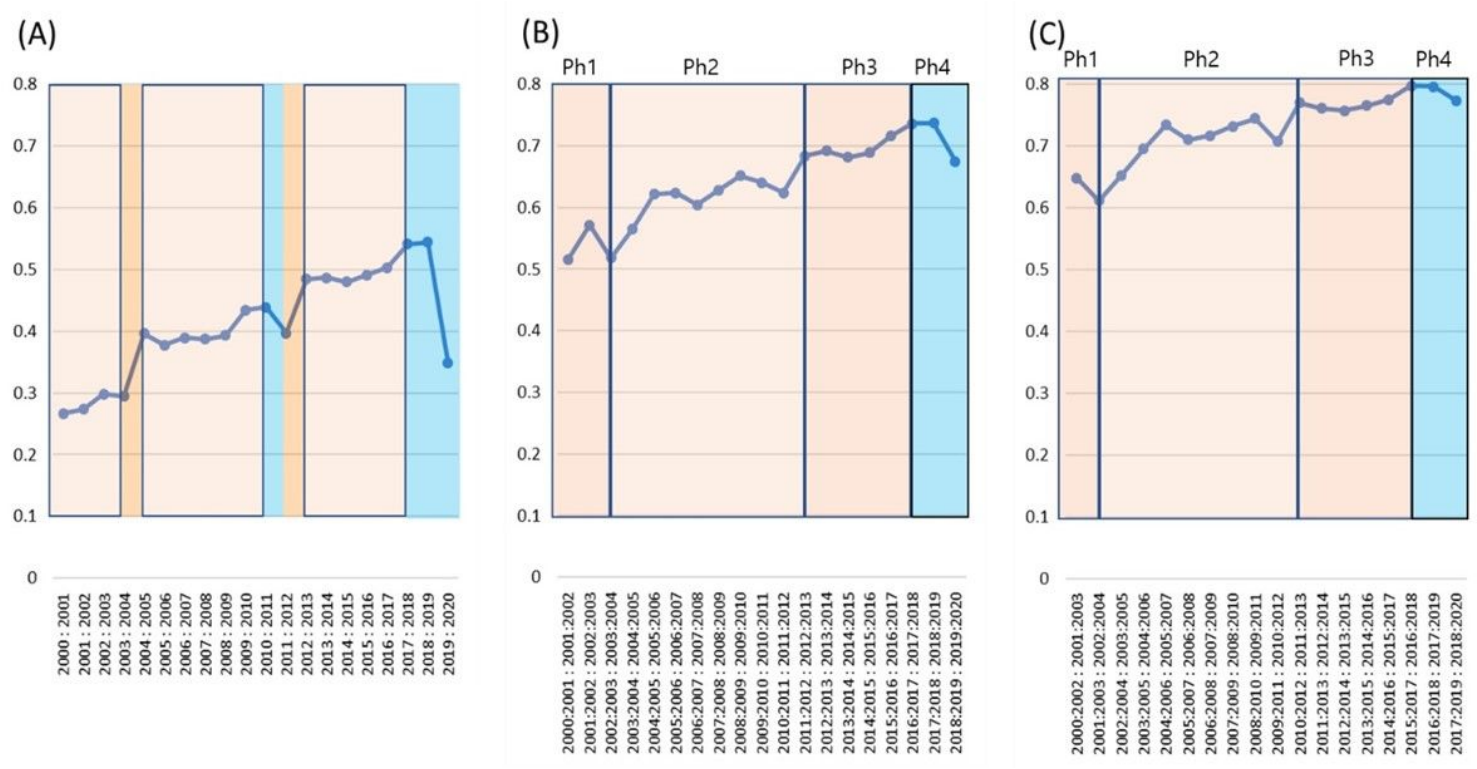

Figure 2

Period analysis. (A) The similarity between 1-year; (B) similarity between a 1-year interval of similarities; (C) similarity between a 2-year interval of similarities

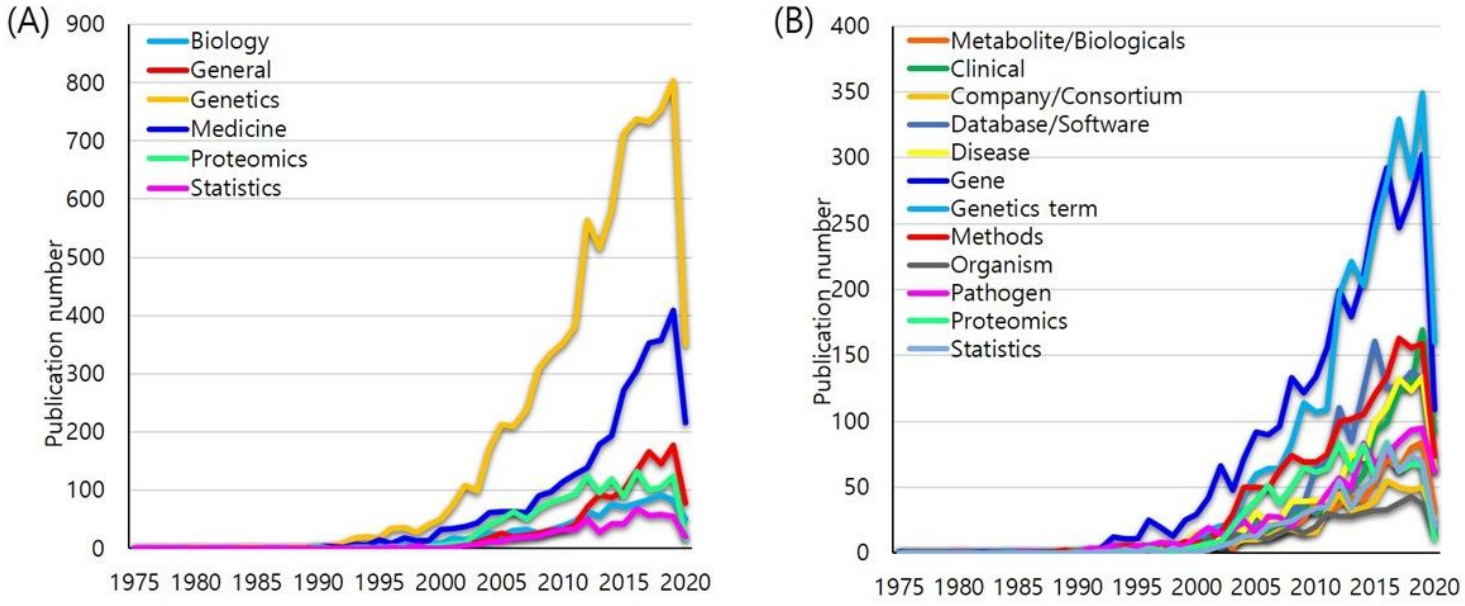

Figure 3

Keyword frequency trend results. (A) publication frequencies' trend of CAT1; (B) publication frequencies' trend of CAT2

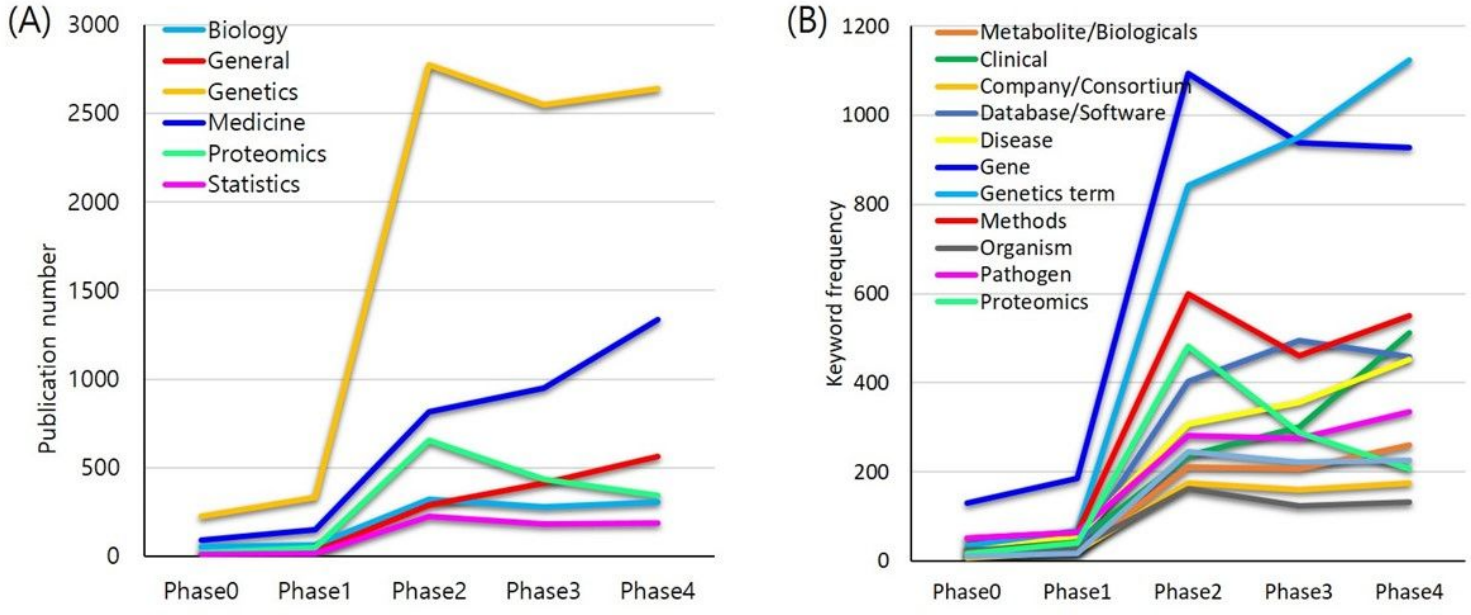

Figure 4

Keyword frequency trend represented by phases. (A) phase trend of frequencies in CAT1; (B) phase trend of frequencies in CAT2

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BMCMedGenSupplementaryInformation.docx