MINI REVIEW

Web resources in post genomic era

Gaurav Jerath and Vibin Ramakrishnan
Department of Biotechnology, Indian Institute of Technology, Guwahati – 781039, India

Correspondence to: vibin@iitg.ernet.in

Abstract

The successful completion of 'Human Genome Project' (HGP) and its availability as an 'Open Source Resource' is arguably the most important advancement in the history of biotechnology and medicinal research in recent times. A number of databases have been reported in the last decade with different aspects of genomic data, including single nucleotide polymorphisms (SNPs), gene expression, protein-protein interactions and more. In this mini-review, we attempt to provide a brief outline of the resources developed after the public domain appearance of the HGP.

1. Human genome project

The Human Genome Project (HGP) was successfully completed by International Human Genome Sequencing consortium in 2003. One of the major principles adopted by the consortium was the public availability of the data generated. Hence, the human genome sequence is freely available in the Genome database maintained by National Center of Biotechnology Information (http://www.ncbi.nlm.nih.gov/guide/genomes-maps/). In the first draft, nearly 30,000 to 40,000 protein coding genes were predicted in the human genome, however, the number was later reduced to less than 21,000 genes.[1] The results also highlighted the identification of more than 1.4 million Single Nucleotide Polymorphisms (SNPs). The completion of this project has also led to the development of several new initiatives. Some of the important initiatives, their availability, and history are presented in this short review.

2. Derivatives of human genome project

The HGP not only opened the floodgates for the analysis of genomic data from different perspectives but also the formation of new directives for a better understanding of its structure and functioning. HGP was soon followed by the development of the International HapMap Project[2], ENCyclopedia of DNA Elements (ENCODE) [3], *etc*. Additionally, ENSEMBL[4] and the UCSC Genome Browser [5] were published to handle the data generated by HGP and analyze it for useful inferences.[6]

2.1. International HapMap project

The project was established in 2002 with an aim to determine the common patterns among the variations in terms of sequence throughout the human genome.[2] Over the years, it has been frequently updated; a second version of the HapMap database which includes more than three million SNPs in four geographically distinct populations are available now. The SNP density of this version is about one per kilo base of sequence and is estimated to contain 25-35 percent of the anticipated 9-10 million SNPs in the human genome.[7]

2.2. ENSEMBL and UCSC genome browser

These databases provide the stable automated annotations for the human genome sequences. ENSEMBL was also a major contributor for deducing the analysis published with the first draft of HGP. The UCSC genome browser has a similar design to ENSEMBL and was primarily designed to support the data generated by HGP. Since their conception, the aim of such data bases has been to deliver the details regarding the sequence in a systematic manner, while also accommodating the framework for data analysis.

2.3. ENCODE project

The ENCyclopedia of DNA Elements Project[3] has an aim of identifying all functional elements, including regions of transcription, transcription factor binding regions and chromatin structures in the human genome. The project was initiated on a pilot scale with a target sequence of 30 megabases in 2004 [3], which was successfully accomplished in 2007 and reported functional elements in this 1% of the genome.[8] Presently ENCODE database contains functional characterization data of about 80 percent of the human genome [9] and can be freely accessed at http://genome.ucsc.edu/encode/.

2.4. The personal genomes project

The concept and necessity of The Personal Genomes Project was presented in 2005 by George Church, he conceived this project as the natural successor of HGP.[10] The project targets the creation of a scientific platform for integration of human genomic, trait and comprehensive environmental data. Such integrated datasets can be deemed essential for the development of functional genomics, providing holistic insights to deduce the underlying mechanisms of human health and diseases.[11] The dataset can be accessed through the web-site http://www.personalgenomes.org.

2.5. GENE ONTOLOGY

The Gene Ontology project was propelled with an aim of developing a platform for structured representation of gene functions and their products in an organism.[12] The genes and products are categorized on the basis of their involvement in a cellular process, molecular function and cellular component to which they belong.[12, 13] A number of tools are available on the web-server, which allows the user to extract ontologies for a list of genes.[12] The utility of such a project is important because information regarding the function of a protein in one organism can lead to essential inferences to its role in other ones. The data can be accessed at: http://www.geneontology.org/. Several other similar resources were published as derivative resources to the HGP, but an extensive discussion is beyond the scope of this mini-review.

3. Primary web-resources

The post genomic era witnessed the development of many useful tools and databases for the analysis and storage of different forms of genomic data. The advent of high-throughput techniques like microarrays has enabled the scientific community to develop new methodologies for gathering information and analyze the data generated by them. We have broadly categorized such resources on the basis of data availability and functionality into three distinct areas.

3.1. Protein-protein interaction databases

The significance of such resources for genomic analysis was clearly outlined in the initial sequence and analysis projections of HGP in important journals.[6, 14] Various cellular processes are governed by molecular interactions of different entities, mostly proteins.[15,16] The development of two-hybrid systems [17] and tandem affinity purification [18] techniques have helped the researchers to record large number of interactions in a single experiment. However, such techniques are also prone to high error rates[19], which demands intervention of specialized computational tools and methodologies.[20]The development of submission guidelines for such resources by the International Molecular Exchange consortium[21], especially, 'Minimum Information' required for reporting a molecular interaction,[22] has greatly influenced the quality of information available in such databases.[23] These resources may be segregated as experimentally and computationally derived interaction resources. The experimentally derived resources include IntAct[23], DIP[24], BioGRID[19], MINT[15], HPRD[25] and MIPS.[26] The computationally derived databases include HomoMINT[27], OPHID[28], PIPs[29], STRING[30] and PrePPI.[31] The details regarding the availability and source of these resources are given in table 1.

Table 1. The list of protein-protein interaction databases

Database	Acronym	URL	Availability
Experimentally Derived			
IntAct	IntAct	http://www.ebi.ac.uk/intact/	Open
Database of Interacting Proteins	DIP	http://dip.doe-mbi.ucla.edu/dip/Main.cgi	Open
Biological General Repository for Interaction Datasets	BioGRID	http://thebiogrid.org	Open
Molecular INTeraction database	MINT	http://mint.bio.uniroma2.it/mint/Welcome.do	Open
Human Protein Reference Database	HPRD	http://www.hprd.org/	
Munich Information Center for Protein Sequences	MIPS	http://mips.helmholtz-muenchen.de/proj/ppi/	Open
Computationally Predicted			
Human Molecular INTeraction database	HomoMINT	mint.bio.uniroma2.it/HomoMINT/Welcome.do	Open
Online Predicted Human Interaction Database	OPHID	http://ophid.utoronto.ca/ophidv2.204/	Open

se			
Human Protein-Protein Interaction Prediction Database	PIPs	http://www.compbio.dundee.ac.uk/www-pips/	Open

Search Tool for the Retrieval of Interacting Genes/ Proteins	STRING	http://string-db.org/	Open

Predicted Protein-Protein Interactions database	PrePPI	http://bhapp.c2b2.columbia.edu/PrePPI/	Open

The IntAct and STRING databases are arguably the most prominent representatives of each category. The rate of data accumulation in these data bases is given in figure 1[23, 32-34] and figure 2[30, 35-39].

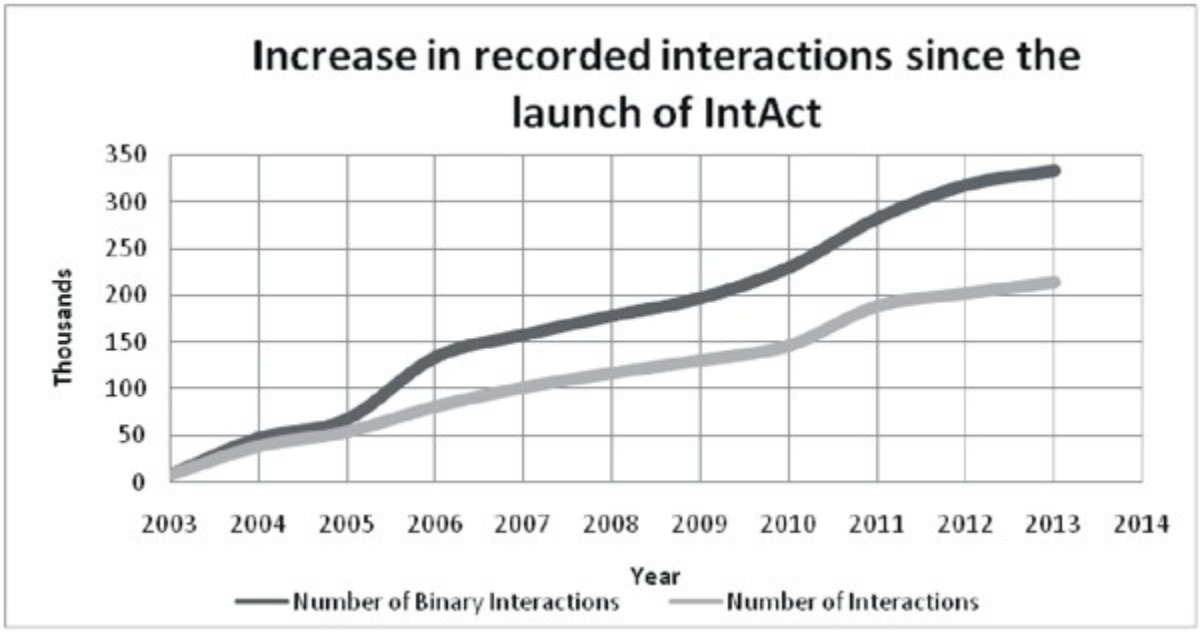

Figure 1. The time-dependent accumulation of data in IntAct database

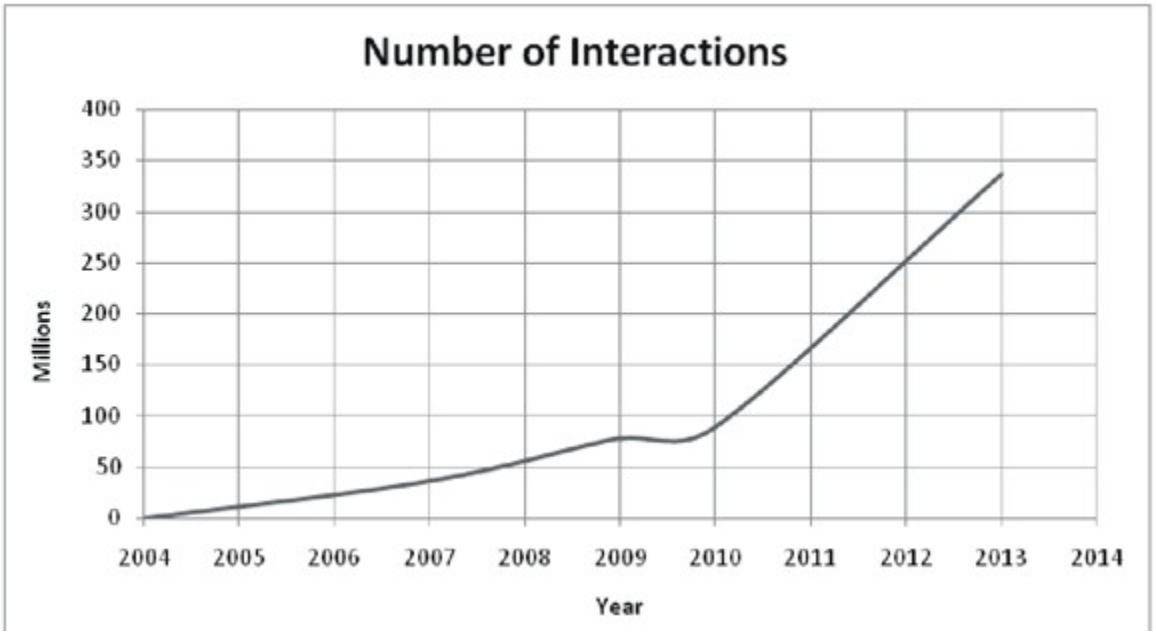

Figure 2. The growth of STRING database since it was published.

3.2. Functional genomics databases

After the successful completion of HGP, the need for complete cataloguing, characterization and understanding of functional elements encoded by the genome became indispensable.[40] Microarrays have been the long time favorites for carrying out such expeditions in functional genomics. The data generated by this array-tiling technique spans the genomic scale and can be interpreted by the use of many open-source tools. The constant use of microarrays in biomedical research led to the 'proposal of Minimal Information About a Microarray Experiment, (MIAME) guidelines[41], which have become a necessity for data submission. The models generated from the analyses of functional genomic data have led to the formulation of a different branch of science, Systems Biology.[42] The utility of functional genomics studies in drug development was highlighted in a recently published study by Lotsch.[43] Numerous resources are available for providing functional data of the genome, but this review covers only major and prominent resources.

Gene Expression Omnibus[44] hosted by NCBI is one of the most important resources for microarray experiments. It was initially designed to meet the requirement of a public repository for gene expression data[45], but later data from other microarray techniques such as chromatin-immunoprecipitation, protein arrays, SNP array, *etc*. was also included. Additionally, GEO provides tools for the analysis of the data stored in its archives.[46] In 2007, a tool named GEOquery was developed[47] that helps in data retrieval from GEO database through Bioconductor hence,

eliminating a number of formatting and parsing problems.

Another major public repository for similar data and analysis is the ArrayExpress[48] database, hosted by European Bioinformatics Institute (EBI). It accepts the submitted data in three forms: arrays, experiments and protocols. Like GEO, it also includes data from other high throughput experiments [49] and as well contains most of the data present in GEO. It is integrated with another database hosted by EBI, Gene Expression Atlas[50], which includes data statistics from meta-analyses of datasets, useful for the inference of condition specific gene expression. Medical University of South Carolina DNA Microarray Database [51] is the other database which has an open-source policy for data sharing. The Stanford Microarray [52] and Genevestigator[53] databases have data dependent policies for sharing data. The addresses and data sharing policies of these databases is indicated in table 2.

Table 2. List of functional genomics databases

Database	Acronym	URL	Availability
Gene Expression Omnibus	GEO	www.ncbi.nlm.nih.gov/geo/	Open
ArrayExpress - functional genomics data	ARRAYEXPRESS	www.ebi.ac.uk/arrayexpress/	Open
Genevestigator	GENEVESTIGATOR	https://www.genevestigator.com/	Free and Paid versions available
Stanford Microarray Database	SMD	http://smd.princeton.edu/	Public(Open) and Private(registered users only) data
Medical University of South Carolina DNA Microarray Database	MUSC DNA Microarray Database	http://proteogenomics.musc.edu/ma/	Open

3.3. Comparative genomics resources/databases

Comparative Genomics studies have been quite useful in the identification of novel oncogenes [54] as well as the identification of biomarkers related to cancer.[55] Some of the resources of this category have already been discussed in previous sections. The ALeleFREquency Database (ALFRED) provides information on frequencies of allelic polymorphisms in various anthropologically defined populations.[56,57] On the other hand, dbSNP (database of Single Nucleotide Polymorphisms), is a database which provides information regarding the nucleic acid sequence variations on a genomic scale.[58] These variations have been majorly classified as single nucleotide variations, simple insertions-deletions, invariant regions and short repeats. Superfamily[59] and Vista Tools[60] are also important resources for analyzing genomic data on a Comparative Genomics front and include a variety of tools for different aspects of data analysis. '1000 genomes' offers an account of variations in human genome in the form of a map.[61] A list of such databases is included in table 3.

Table 3. List of comparative genomics tools

Database/Tools	Acronym	URL	Availability
UCSC Genome Browser	None	http://genome.ucsc.edu/	Open
ENSEMBL Genome Browser	ENSEMBL	www.ensembl.org/	Open
HapMap	HAPMAP	http://hapmap.ncbi.nlm.nih.gov/	Open
Superfamily	SUPERFAMILY	http://supfam.cs.bris.ac.uk/SUPERFAMILY/	Open
Vista-tools for Comaprative Genomics	VISTA	http://genome.lbl.gov/vista/index.shtml	Open
1000 Genomes	1000Genomes	http://www.1000genomes.org/	Open

Disease specific genomic resources

Many databases have been specifically developed for different diseases, with an intent to provide detailed information for particular diseases. The following section outlines the databases and tools developed specifically for cancer and cardiovascular diseases.

3.4. Cancer specific databases

Cancer is an epigenetic disease characterized by uncontrolled cell proliferation.[62] Cancer is also one of the leading causes of deaths across the globe. Many databases have been developed for cancer, though many of them are very specific to a particular type of cancer. The databases include both functional and comparative genomics data with the common goal of discovering an effective treatment strategy for this fatal disease. The functional databases include ONCOMINE [63] and caArray while the comparative genomics databases comprise of Can Gene Base[64], COSMIC[65]and Cancer Genome Atlas.[66] Utilities of these databases are detailed in table 4.

Table 4. List of cancer specific data bases

Database	Acronym	URL	Availability
Oncomine	None	https://www.oncomine.org	Academic and Commercial Versions
caArray	None	https://cabig-stage.nci.nih.gov/community/tools/caArray	Open
Cancer Gene Database Development	CanGeneBase	http://bioinfo.au-kbc.org.in/cancerdb/	Open
The Cancer Genome Atlas	None	http://cancergenome.nih.gov/	Open
Catalogue of somatic mutations in cancer	COSMIC	http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/	Open

3.5. Databases specific to cardiovascular diseases

Cardiovascular diseases include a fairly long list of diseases related to heart and vascular system like congenital heart disease, coronary artery disease, atherosclerosis, ischemic heart disease, *etc*. Collectively, cardiovascular diseases are the leading causes of death worldwide, for a very long time. The databases developed for this collection of diseases include genomic as well as surgical data. The Gene Ontology project has also taken a dedicated initiative for the public availability of functional annotation of the genes involved in the function of heart development and cardiovascular diseases. A limited list of such databases is given in table 5.

Table 5. List of data bases for cardiovascular diseases

Database	Acronym	URL
Coronary Artery Disease Gene Database	CADgene	http://www.bioguo.org/CADgene/
CardioGenomics	None	http://cardiogenomics.med.harvard.edu/home
CATHGEN	None	http://cathgen.duhs.duke.edu
GENetics of Early onset CARdiovascular Disease	GENECARD	http://www.chg.duke.edu/diseases/genecard.html
The Cardio Research Web Project	Cardio-Research	http://www.cardio-research.com/databases

4. Summary

The conclusion of the Human Genome Project has greatly influenced the development of functional and comparative genomics. The availability of high throughput data in the public domain has led to the analysis of this data from multiple perspectives, targeted to meet specific objectives. Overall a vigorous effort from various scientific groups in fulfilling the designed objectives of making human genome project data public seems to have met. At least, it gives a hope that affordable and more personalized healthcare solution can be a realistic dream in the

following decade. However, many parts of the 'complete picture' are still missing, and the lack of proper tools for analysis of such data has marred many important conclusions. The development of new techniques for data generation needs to be supplemented with proper strategies for its analysis. The advent of specialized databases has helped the researchers working in those areas in getting processed information from a single source. The main challenge ahead is the compilation of data from various types of databases on a single platform for meaningful inferences.

References

1. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences* 2007; **104**:19428-19433.

2. The International HapMap Consortium. The International HapMap Project. *Nature* 2003; 426:789-796.

3. The ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004; **306**:636-640.

4. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, *et al.* Ensembl 2013. *Nucleic Acids Research* 2013; **41**:D48-D55.

5. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, *et al.* The Human Genome Browser at UCSC. *Genome Research* 2002; **12**:996-1006.

6. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 2001; **409**:860-921.

7. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**:851-861.

8. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; **447**:799-816.

9. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012; **489**:57-74.

10. Church GM. The Personal Genome Project. *Mol Syst Biol* 2005; 1.

11. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, Wu X, *et al.* A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences* 2012; **109**:11920-11927.

12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet* 2000; **25**:25-29.

13. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 2004; **32**:D258-D261.

14. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al.* The Sequence of the Human Genome. *Science* 2001; **291**:1304-1351.

15. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research* 2012; **40**:D857-D861.

16. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* 2002; **30**:303-305.

17. Legrain P, Selig L. Genome-wide protein interaction maps using two-hybrid systems. *FEBS Letters* 2000; **480**:32-36.

18. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotech* 1999; **17**:1030-1032.

19. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon G, Myers C, *et al.* Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *Journal of Biology* 2006; **5**:11.

20. Scott M, Barton G. Probabilistic prediction and ranking of human protein-protein interactions. *BMC Bioinformatics* 2007; **8**:239.

21. Orchard S, Kerrien S, Jones P, Ceol A, Chatr-aryamontri A, Salwinski L, *et al.* Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics* 2007; **7** Suppl 1:28-34.

22. Orchard S, Salwinski L, Kerrien S, Montecchi-Palazzi L, Oesterheld M, Stumpflen V, *et al.* The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol.* 2007; **25**:894-898.

23. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, *et al.* The IntAct molecular interaction database in 2010. *Nucleic Acids Research* 2010; **38**:D525-D531.

24. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research* 2004; **32**:D449-D451.

25. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, *et al.* Human Protein Reference Database-2009 update. *Nucleic Acids Research* 2009; **37**:D767-D772.

26. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, *et al.* MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Research* 2004; **32**:D41-D44.

27. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics* 2005; **6**:S21.

28. Brown KR, Jurisica I. Online Predicted Human Interaction Database. *Bioinformatics* 2005; **21**:2076-2082.

29. McDowall MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. *Nucleic Acids Research* 2009; **37**:D651-D656.

30. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* 2013; **41**:D808-D815.

31. Zhang QC, Petrey D, Garzon JI, Deng L, Honig B. PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Research* 2013; **41**:D828-D833.

32. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Research* 2004; **32**:D452-D455.

33. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, *et al.* IntAct-open source resource for molecular interaction data. *Nucleic Acids Research* 2007; **35**:D561-D565.

34. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Research* 2012; **40**:D841-D846.

35. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, *et al.* STRING 8-a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Research* 2009; **37**:D412-D416.

36. Mering Cv, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* 2003; **31**:258-261.

37. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* 2011; **39**:D561-D568.

38. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, *et al.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research* 2005; **33**:D433-D437.

39. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, *et al.* STRING 7-recent developments in the integration and prediction of protein interactions. *Nucleic Acids Research* 2007; **35**:D358-D362.

40. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature* 2003; **422**:835-847.

41. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C *et al.*

Minimum information about a microarray experiment (MIAME)[mdash]toward standards for microarray data. *Nat Genet* 2001; **29**:365-371.

42. Auffray C, Imbeaud S, Roux-Rouquie M, Hood L. From functional genomics to systems biology: concepts and practices. *Comptes Rendus Biologies* 2003; 326:879-892.

43. Lotsch J, Doehring A, Mogil JS, Arndt T, Geisslinger G, Ultsch A. Functional genomics of pain in analgesic drug development and therapy. *Pharmacology & Therapeutics* 2013; 139:60-70.

44. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, *et al.* NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Research* 2013; **41**:D991-D995.

45. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 2002; **30**:207-210.

46. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, *et al.* NCBI GEO: mining tens of millions of expression profiles-database and tools update. *Nucleic Acids Research* 2007; **35**:D760-D765.

47. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 2007; **23**:1846-1847.

48. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, *et al.* ArrayExpress update-trends in database growth and links to data analysis tools. *Nucleic Acids Research* 2013; **41**:D987-D990.

49. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, *et al.* ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* 2003; **31**:68-71.

50. Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, *et al.* Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 2012; **40**:D1077-D1081.

51. Argraves GL, Barth JL, Argraves WS. The MUSC DNA Microarray Database. *Bioinformatics* 2003; **19**:2473-2474.

52. Hubble J, Demeter J, Jin H, Mao M, Nitzberg M, Reddy TBK, *et al.* Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Research* 2009; **37**:D898-D901.

53. Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, *et al.* Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes. *Advances in Bioinformatics* 2008; **2008**:5.

54. Li N, Kaur S, Greshock J, Lassus H, Zhong X, Wang Y, *et al.* A Combined Array-Based Comparative Genomic Hybridization and Functional Library Screening Approach Identifies mir-30d As an Oncomir in Cancer. *Cancer Research* 2012; **72**:154-164.

55. Huang CC, Jeng JY, Tu SH, Lien HH, Huang CS, Lai LC, *et al.* A preliminary study of concurrent gains and losses across gene expression profiles and comparative genomic hybridization in Taiwanese breast cancer patients. *Translational Cancer Research 2013;2:1 (February)*.

56. Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, *et al.* ALFRED: the ALelle FREquency Database. Update. *Nucleic Acids Research* 2003; **31**:270-271.

57. Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK. ALFRED: An allele frequency database for anthropology. *American Journal of Physical Anthropology* 2002; **119**:77-83.

58. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 2001; **29**:308-311.

59. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, *et al.* SUPERFAMILY-sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* 2009; **37**:D380-D386.

60. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Research* 2004; **32**:W273-W279.

61. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**:56-65.

62. Hanahan D, Weinberg R. Hallmarks of cancer: The next generation. *Cell* 2011;**144**:646-674.

63. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB, *et al.* Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia.* 2007; **9**:166-180.

64. Kumar GR, Subazini TK, Subha K, Rajadurai CP, Prabakar L. CanGeneBase (CGB)--a database on cancer related genes. *Bioinformation.* 2009; **3**:422-424.

65. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research* 2011; **39**:D945-D950.

66. Robbins DE, Gruneberg A, Deus HF, Tanik MM, Almeida JS. A self-updating road map of The Cancer Genome Atlas. *Bioinformatics* 2013;**29**:1333-40.