Internet databases and resources for cytogenetics and cytogenomics

Etienne De Braekeleer, Jean Loup Huret, Hossain Mossafa, Katriina Hautaviita, Philippe Dessen

Haematological Cancer Genetics & Stem Cell Genetics, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, United Kingdom; Medical Genetics, Dept Medical Information, University Hospital, F-86021 Poitiers, France; Laboratoire CERBA, 95310 Saint Ouen l’Aumone, France; (Mouse genomics, Wellcome Trust Sanger Institute); UMR 1170 INSERM, Gustave Roussy, 114 rue Edouard Vaillant, F-94805 Villejuif, France.

Published in Atlas Database: April 2016
DOI: 10.4267/2042/62675

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 2.0 France Licence. © 2017 Atlas of Genetics and Cytogenetics in Oncology and Haematology

Abstract

Databases devoted stricto sensu to cancer cytogenetics are the "Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer" (http://cgap.nci.nih.gov/Chromosomes/Mitelman), the "Atlas of Genetics and Cytogenetics in Oncology and Haematology" (http://atlasgeneticsoncology.org), and COSMIC (http://cancer.sanger.ac.uk/cosmic). However, cancer being a complex multi-step process, cytogenetics are broaden to "cytogenomics", with complementary resources, including resources on proteins and cancer. These resources are essential to both practical and theoretical knowledge in cytogenomics of cancer. Must be briefly reviewed: general databases (nucleic acid and protein sequences databases and bibliographic ones), cancer genomic portals associated to recent international integrated programs, such as TCGA or ICGC, fusion genes databases, genomic sequences and transcripts databases (with different cartography browsers), array CGH databases and structural variation databases for copy number, polymorphisms and mutation databases, databases on proteins (structure and function with implication of mutations and rearrangements), databases on diseases, databases and books on pathology, cancers, and patient associations and interfaces between science and patients. Other resources such as the International System for Human Cytogenetic Nomenclature (ISCN), the International Classification of Diseases for Oncology (ICD-O), the Human Gene Nomenclature Database (HGNC), and the Nomenclature for the description of sequence variations allow a common language. Data within the scientific/medical community should be freely available. However, most of the institutional stakeholders are now gradually disengaging, and well known databases are forced to beg or to disappear (which may happen!).

Keywords: Cytogenetic, Cancer, Database, Mitelman Database, Atlas of Genetics and Cytogenetics in Oncology and Haematology, COSMIC, PubMed, GenBank, TCGA, ICGC, UniProt, OMIM, IARC, ISCN, ICD-O, HGNC.

INTRODUCTION

In each cancer case there is a genetic event present (Stratton MR et al., 2009). Cytogenetics has been a major player in understanding genetics behind cancer, providing specific keys for diagnosis and prognostic assessments, as well as enabling the subclassification of otherwise seemingly identical disease entities (Mertens F et al., 2015). This "Deep Insight is dedicated to cytogenetics resources will highlight the various facets in theoretical understanding of cancer and the consequent practical strategies in treating the disease.

Brief history
In 1914, Boveri stated that the heritable acquired characteristics of cancer cells are brought about by a disturbance of the normal chromosomal balance (Boveri, 1914). This theory was supported by a wealth of experimental data showing that cancer originates in a single cell through acquired genetic changes. The investigation in the 1950s, on ascites tumors that were induced experimentally or observed in patients tended to confirm that cytogenetic aberrations are an important and integral part of tumor development and evolution. These cytogenetic studies demonstrated that certain laws could direct neoplasia-associated chromosomal variability. Like selective pressures, where any changes in the surrounding tumor would modify the equilibrium, causing a change where the most viable chromosomal profile is prevailing in the new environment.

The importance of cytogenetics boomed since the discovery of the first chromosomal anomaly reported by Peter Nowell and David Hungerford in 1960, linking the Philadelphia (Ph) chromosome to chronic myeloid leukaemia (CML) (Nowell and Hungerford, 1960). It was the first assessment in detecting chromosomal anomaly in human leukaemia and seemed reasonable that it was the cause of origin for CML. This discovery was the first strong assessment to Boveris theory. This observation stimulated the field to find other karyotypic anomalies in other cancers. Unfortunately, a heterogenous panel of chromosomal rearrangements was detected in what seemed to be the same cancer. This was a terrible setback for the arguments stating karyotypic anomalies as the origin of cancer. The explanation was that chromosomal rearrangements were an epiphenomenon that could appear during tumor progression without having any pathogenetic consequences.

In the 1970's the situation changed dramatically when chromosomal banding techniques invented by Caspersson and Zech (Caspersson T et al., 1970) were introduced. This process gave an option to identify individual chromosomes, which were defined by a unique banding pattern. The description of chromosomal rearrangements immediately became clearer providing more gravity to the conclusions drawn. This was a new era for cancer cytogenetics showing an increase in the numbers of aberrant human malignant and benign karyotypes. In the 1980s, the onset of molecular genetics techniques intensely widened our understanding of the pathogenetic progression underlying the neoplastic process. These techniques provided an opportunity to characterise the chromosomal breakpoints at the molecular level and has highlighted two classes of genes implicated in these karyotypical rearrangements: the oncogenes and the tumor suppressing genes.

MYC and BCR/ABL1: One of the first oncogenes described as activated by chromosomal rearrangement is MYC, which was characterised in Burkitt lymphoma studies. Another example is the translocation between ABL1 and BCR. Peter Nowell and David Hungerford first described a recurrent presence of an extra chromosome in CML patients in 1960 (Nowell PCH and Hungerford DA, 1960). In 1973 Janet D. Rowley used quinacrine coloration to prove this chromosome to be the result of a translocation between chromosomes 9 and 22 (Rowley JD, 1973a). Only as late as 1982 de Klein et al. showed that the genes ABL1 and BCR were fused together giving rise to an abnormal gene (de Klein A et al., 1982). With these new techniques, each chromosome and chromosome region could be identified on the basis of their unique banding pattern, giving daylight to previously undetectable subtle rearrangements. By this technique she identified the recurrent translocation t(8;21)(q22;q22) (Rowley JD, 1973b). These findings evoked interest in the cytogenetic analysis of other haematological malignancies. The number of reported balanced rearrangements has increased, in particular translocations including t(8;14)(q24;q32), t(2;8)(p11;q24) and t(8;22)(q24;q11) in Burkitt lymphoma (Zech L et al., 1976 ; Berger R et al., 1979 ; Miyoshi I et al., 1979 ; Van Den Berghe H et al., 1979 ), t(4;11)(q21;q23) in acute lymphoblastic leukaemia (ALL) (Oshimura M et al., 1977 ) t(15;17)(q22;q21) in acute promyelocytic leukaemia (APL) (Rowley JD et al., 1977), and t(14;18)(q32;q21) in follicular lymphoma (Rowley JD et al., 1977). During this fruitful period the first specific translocation in an animal model was found, a mouse plasmacytoma, which is a B cell malignancy displaying similar characteristics to human Burkitt lymphomas (Ohno S et al., 1979).

The following decade witnessed a rise in number of results from malignant solid tumors, mainly sarcomas but also a few carcinomas. Several of the aberrations identified were as specific as the ones that were previously described in haematological cancer: t(2;13)(q36;q14) in alveolar rhabdomyosarcoma (ARMS) (Seidal T et al., 1982), t(11;22)(q24;q12) in Ewing sarcoma (Aurais A et al., 1983 ; Ture-Carel C et al., 1983), t(X;1)(p11;q21) in Kidney cancer (de Jong B et al., 1986) and t(6;9)(q23;p23) in alivary gland tumors (ACC) of the salivary glands (Stenman G et al., 1986). Evidence was showing that many benign tumors were bearing characteristic rearrangements, including reciprocal translocations such as t(3;8)(p21;q12) in salivary gland adenoma (SGA) (Mark J et al., 1980) and t(3;12)(q27;q13) in lipoma (Heim S et al., 1986 ; Ture-Carel C et al., 1986). Although the vast majority of fusion genes are formed by balanced translocations, they can also be produced by interstitial deletions. These were first
identified in the 1990s, amongst them the fusion between genes STIL (STIL/TAL1 interrupting locus) in T-ALL (Bernard O et al., 1991). Since then, many others were observed with more or less extensive deletions, duplications and/or amplifications in the breakpoint regions (Barr FG et al., 1996; Simon MP et al., 1997; Sinclair PB et al., 2000; Müller E et al., 2011). Gene fusion can also arise from copy number shifts like in the aforementioned fusion gene USP16/RUNX1 (ubiquitin specific peptidase 16 and runt related transcription factor 1) in chronic myelomonocytic leukemia (Gelsi-Boyer V et al., 2008) and in the fusion gene SET/NUP214 (SET nuclear proto-oncogene and nuclear pore complex protein Nup214) in T-ALL (Van Vlierberghe P et al., 2008; Mullighan CG et al., 2009; Santo EE et al., 2012; Plaszczyca A et al., 2014) (Figure 1).

Technical developments
In the late 1970s, various technical developments helped in solving what molecular consequences the oncogenic chromosomal rearrangements could have. These techniques enabled the identification and characterisation of genes that were located at the breakpoints of chromosomal rearrangements. The genes implicated in MPC, Burkitt lymphoma and CML proved to be pivotal for the comprehension of the mechanism underlying chromosomal rearrangements. The engineering of fluorescence in-situ hybridization (FISH) enabled several chromosomal structures to be identified simultaneously. This significantly improved the location of breakpoints on chromosomes. It also considerably reduced the scale of which chromosomes could be observed and broadened the type of rearrangements that could be observed (cryptic rearrangements). The big advantage of the FISH technique is that it can also be used for non-dividing cells (interphase nuclei). FISH probes of a specific gene can identify new partner genes, like in the case of mixed lineage leukemia (MLL, KMT2A) gene (De Braekeleer E et al., 2009; Meyer C et al., 2013).

Although cytogenetic analyses are unquestionably crucial for the identification of fusion genes and rearrangements, there are certain limits to this technique. Firstly, revealing chromosome bandings requires having access to in-vitro living, dividing cells so that metaphases can be observed. Secondly, some tumor types can have very complex genomes which makes it difficult to understand the full story and distinguish the primary aberrations and origin of the cancer development from the bulk of the rearrangements (Speicher MR and Carter NP, 2005).

In the 1990s, the progress of high throughput tools for global genetic analyses, such as array based platforms for gene expression and copy number profiling, gave rise to new methods for observing chromosomal rearrangements. These techniques were not ideal either since balanced chromosomal rearrangements could pass undetected or the analysis of expression profiles could prove to be tricky. On
Internet databases and resources for cytogenetics and cytogenomics

De Braekeleer E, et al.

the other hand, they presented a higher level of resolution than in chromosome banding and didn't require prior cell culturing (Pinkel D and Albertson DG, 2005; De Braekeleer E et al., 2014). The first novel gene fusion detected with the analysis of gene expression pattern of a tumor was the fusion of the transcription factor PAX3 gene with the nuclear receptor co-activator 1, NCOA1 gene. By focusing on genes presenting outlined values of expression, the fusions genes implicating the transmembrane protease serine 2 gene (TMPRSS2) with two genes encoding ETS transcription factors. The first is v-ets avian erythroblastosis virus E26 oncogene homolog (ERG) and the second is ets variant 1 (ETV1) (Tomlins SA et al., 2005). It was the first report of specific fusion genes implicated and representing a major subset of a common epithelial malignancy. By using a modification of this method, other fusion genes were discovered in many tumor types, such as tenosynovial giant cell tumor, lung cancer and chondrosarcoma (West RB et al., 2006; Rikova K et al., 2007; Soda M et al., 2007; Wang L et al., 2012).

The introduction of deep sequencing technologies a few years ago gave a new insight to identify new fusions genes either at DNA or RNA level. The combination of detailed information (base pair level) and broad (genome-wide) on DNA, transcriptome, structural variants and fusion transcripts could be obtained without any prior information on the cytogenetic features of the cancer cells. The initial study using deep sequencing to detect fusion genes or chromosomal rearrangements were done on established cell lines (Campbell PJ et al., 2008). The analysis of primary samples from common cancer (Maher CA et al., 2009a; Maher CA et al., 2009b), such as carcinomas of the breast (Stephens PJ et al., 2009), colon (Cancer Genome Atlas Research Network, 2013), lung (Cancer Genome Atlas Research Network, 2012), prostate (Cancer Genome Atlas Research Network, 2014), uterus (Cancer Genome Atlas Research Network, 2013) as well as leukaemias and lymphomas (Steidl C et al., 2011; Welch JS et al., 2011; Roberts KG et al., 2012), came afterward. One study draws a bridge between over several several cancers by cumulating the bioinformatics data of 4,366 cancers from 13 different tumor types that were previously studied within the Cancer Genome Atlas (TCGA) network. The outcome was the description of 8,600 different fusion transcripts (Yoshihara K et al., 2015). These results have dramatically changed the gene fusion landscape with the identification of more than 10,000 fusion genes with more than 90% of these having been identified by various deep-sequencing approaches during the last 5 years (Mitelman F et al., 2016; Huret JL et al., 2013).

The high resolution of deep sequencing gave the possibility to identify the vast majority of genes implicated in chromosomal rearrangements that would have been complicated or impossible to identify by conventional cytogenetic techniques. Indeed, 75% of the genes fusions first detected by deep sequencing are intrachromosomal and approximately 50% are between genes located in the same chromosome band (Mitelman F et al., 2016). Large majority of genes, - already described in the literature before the deep sequencing era- were embedded in extensive networks like MLL in leukaemias, EWS RNA-binding protein 1 (EWSR1) in sarcomas and rearranged during Transfection Protooncogene (RET) in carcinomas (Mitelman F et al., 2007). However, this picture has somewhat changed with the massive increase of fusion genes that were added with genome-wide studies. The fact that these studies were mainly focusing on previously uncharacterized tumor types brought a lot of new networks emerging from rarer gene fusions than leukaemias, lymphomas and sarcomas. Furthermore, carcinomas often show highly rearranged genomes, with numerous mutations at the gene and chromosome levels and it may be that the genes detected by deep sequencing are the results of chance events caused by chromosomal instability, as vast majority of fusion transcripts were associated with amplification or deletion events at the DNA level (Yoshihara K et al., 2015; Mitelman H et al., 2016; Huret JL et al., 2013; Mitelman F et al., 2007; Kalyana-Sundaram S et al., 2012). Transcription-induced gene fusion (TIGF) or Trans-TIGF, when they happen on different chromosomes, results in the fusion of transcripts from non-adjacent genes without a corresponding fusion at DNA level (Gingeras TR, 2009; Rickman DS et al., 2009; Meyer C et al., 2009; Hedegaard J et al., 2014). Certain have been shown to have no impact, since they were expressed in normal tissues like the fusion genes JAZF zinc finger 1 (JAZF1)/SUZ12 a polycomb repressive complex 2 subunit and PAX3/ . Others, implicating the gene MLL, to be the driving mutation (Meyer C et al., 2009).

The prognostic and treatment value of chromosomal rearrangements and mutated genes: The high correlation between recurrent gene fusions and tumor subtypes has made them the ideal maker for diagnostic purposes. This correlation is also important in treatment stratification, the best example being the different fusion of MLL in AML (Meyer C et al., 2009). The routine molecular strategy to detect these fusion genes is the use of cytogenetics, FISH, RT-PCR and deep sequencing. The mounting knowledge of the clinical importance of gene fusions, as well as various chromosomal rearrangements, has gradually led to an increasing emphasis on genetic features in the classification of tumors. The latest World Health Organisation (WHO) classification, translocation and/or gene fusion status is mandatory for the diagnosis of some types of tumors, such as "AML with
t(8;21)(q22;q22), RUNX1/RUNX1T1" and "B lymphoblastic leukaemia/lymphoma with t(5;14)(q31;q32), IL3/IGH". For other cancers, such as alveolar soft part sarcoma and synovial sarcoma, it is considered as a distinctive defining element of the neoplasm (Fletcher CD, 2014; Swerdlow SH et al., 2016). Since fusion genes are diagnostic markers, they can also be used as markers for monitoring minimal residual disease following treatment (De Braekeleer E et al., 2014; Ommen HB and Hokland P, 2011). Currently, this strategy is in clinical use mainly for haematological disorders but the improvements in the detection and enrichment of circulating cancer cells and DNA suggest that solid tumors with gene fusions might also be monitored in a similar way (Crowley E et al., 2013; Karabacak NM et al., 2014; Watanabe M et al., 2014; Yu KH et al., 2014; Baccelli I et al., 2013).

It is important to mention that the detection of the fusion gene can be used to monitor the progression or the relapse of the cancerous cells but it doesn't need to be an important actor in the neoplastic phenomenon, as long as they are representative and specific of the neoplastic cells (Leary RJ et al., 2010).

Research on fusion genes paved the way to develop specific drugs targeting chimeric proteins. The tyrosine kinase inhibitor Imatinib, approved in 2001, was the first drug specifically designed to target the chimeric protein BCR/ABL1 in CML (Druker BJ et al., 2001; Druker BJ et al., 2001) by blocking its kinase activity. This drug dramatically improved the lifespan and life quality of patients bearing CML. The immense success of imatinib spurred interest in developing new compounds against the chimeric proteins, all of which are kinase inhibitors. Different tumors have shown to display various fusions involving kinase-encoding genes, such as ALK, BRAF, Fibroblast growth factor receptor 3 (FGFR3), neurotrophic tyrosine kinase receptor type 1 (NTRK1), RET and ROS1 (Yoshihara K et al., 2015; Fletcher CD, 2014; Swerdlow SH et al., 2016). These fusion genes are occurring at low frequencies but if merged they represent a considerable number of patients. Stratification strategies considering the genotype and phenotype of the tumor would contribute greatly to identifying patients with these very promising treatment targets. Many new compounds are currently being tested in clinical models although others have reached the Phase 1 and Phase 2 stages in clinical trials, for example, chromatin modifier such as MLL (Malik R et al., 2015), DOT1L (Chen CW et al., 2015), BRD4 (Dawson MA et al., 2011) or EZH2 (McCabe MT et al., 2012; Fillmore CM et al., 2015).

The need for organising data banks
Since discovering their involvement in cancer initiation, progression and evolution, chromosomal rearrangements have triggered wide, increasing interest to understanding them better. The amount of genes involved has increased, the network underlying certain genes has been resolved and the mechanistic aspect is unravelled. Unfortunately, a lot of work has to be done before cancer has been eradicated. One of the steps is to synthesise all the information and make it available in order to increase the common knowledge of genes that are implicated and their interactions with other pathways in the cell. The importance of creating data banks and reporting various chromosomal rearrangements has been recognized since the 80's.

1981: Human Genome Mapping
The information on chromosome modification in cancer has been included as part of the Human Genome Mapping (HGM) workshop since 1981. The provision of up-to-date information of all chromosomal rearrangements was the initial goal. This means including all case reports, which are suspected to be the starting point of tumor development or a contribution to the proliferation but also complex karyotypes with several cytogenetic anomalies or secondary modifications leading to the evolution and resistance to treatment. The increasing number of cases, reports and the multitude of cytogenetically abnormal neoplasms made it too challenging to include everything in the database. In 1991, the HGM decided to focus only on aberrations repeatedly found as sole anomalies in a few given tumor types. As a consequence the number of recurrent changes was severely underestimated, especially in solid tumors where single anomalies are a rare finding. This illustration of chromosomal anomalies mainly represents the tip of the iceberg since the generalisation and improvements in classic cytogenetic techniques and the development of new techniques have considerably increased the number of reports of chromosomal rearrangements in different types of tumors. Several of these anomalies may be of diagnostic and prognostic importance, as well as a large amount of details of molecular analysis.

1983: Catalog of Chromosome Aberrations in Cancer
In 1983, Felix Mitelman published a colossal manuscript that was a supplement to Cytogenetics and Cell Genetics. The goal of this publication was to catalogue all known chromosomal rearrangements. The complexity of the data pushed the laboratories and institutes to adopt computerised methods to compile, revise and index the information. Many cytogeneticist, clinicians and cell biologists were in the demand for a systematic, concise and uniform presentation of material. The vast body of literature was making it complicated to evaluate if a chromosomal abnormality had been described before or not. To facilitate this process,
Mitelman presented a compilation of 3,844 published and unpublished cases from colleagues or from his own laboratory. The two volumes presented all the implicated genes, chromosomes and rearrangements known. This set of two books was the first of its kind but far from the last. For several years these two volumes accompanied the bookshelves of several cytogenetists and oncologists. A re-edition of this work took place in 1985 with data of new cases and improved data of cases already described. The number of cases had now increased to a bit more than 5,000. The number of cases increased with each edition so that by the fifth edition it was composed of two large volumes of more than 4,000 pages, making it arduous to use. The sixth edition had already more than 30,000 cases in it. To make it more user friendly it was then published as a CD. The number of cases would still continue to increase and this information was not freely available. Felix Mitelman then had the idea to display the information on the Internet, rendering it freely available. In 2000, the catalogue became accessible for the public under the name Mitelman Database of Chromosome Aberrations in Cancer associated to the Cancer Genome Anatomy Project internet site and under the supervision of the National Cancer Institute (see below).

1997: Atlas of Genetics and Cytogenetics in Oncology and Haematology
How did the idea of the Atlas come about? Prognosis for leukaemia depends on the genes involved: 5 years survival rate: 6% in the inv(3)(q21q26) RPN1/MECOM leukemia, 100% in the dic(9;12)(p13;p13) PAX5/ETV6 leukemia. Treatment depends on the severity of the disease. However, thousands of genes were discovered to be implicated in cancer (14,000 unique fusion transcripts have been detected), and 1,200 types of solid tumors exist. Some cancers are frequent while many others are very rare (many with only 1 published case). This is particularly true for leukemia subtypes of which there are more than 1,000! 25,000 new publications concerning human cancer genetics are added each year to PubMed. No-one has the whole required knowledge, necessary to guide the treatment procedure in case of a rare disease. The following conclusion was made that huge databases were required to collect and summarize data on these rare diseases in order to produce meta-analyses. The Atlas has been established for that reason; to contribute to 'metomedicine', meaning the mediation between the knowledge and the knowledge users in medicine.

Besides resources dedicated only to cytogenetics, a quick overview of resources in surrounding areas "Cancer Cytogenetics", stricto sensu, deals with chromosomes and cancer. "Cytogenetics" means "Cell Genetics" ("cyto" comes from κύτος, in the meaning of the term "the cell"); "Cytogenomics", as coined by Alain Bernheim, (Bernheim A et al., 2004) (from a princeps paper in French in 1998), means the "genetics -as a whole- of the cell", with complex interconnections and interactions between these operators. As is known for long, "one-gene-one-reaction" (Beadle GW, 1945) (understood today as "one-gene-one-protein"), and we can infer from "Cyto-genomics" to the terms "Cyto-transcriptome" and "Cyto-proteomics", or, in a more holistic approach, (and more simply) "Cell Biology".

Cancer is now known as being a multi-step process, with genetic events at almost each step. Therefore, the "Cancer Cytogenetics" research field should incorporate knowledge of the "Cell Biology" of normal and cancerous cells, gene fusions, mutations or copy number variation, epigenetics, protein domains, metabolic or signaling pathways, as well as consequences of these cytogenomic rearrangements and disorders in the pathogenesis of cancer, from gross and microscopic pathological presentation to patients and diseases, clinical pictures, and, even, to epidemiological data given by cancer registries.

It is useful for the cancer cytogeneticists to have an easy and quick access to databases and books of these surrounding subject areas. Therefore, besides resources of cancer cytogenetics, we will mention other resources, including resources on proteins and resources on cancer.

Presently, Internet provides access to a vast and complex network of knowledge that can make it challenging for you to find the answer to your questions. Several databases are freely accessible, but unfortunately not all of them are user friendly. We will briefly describe the main resources in the following pages.

Recent reviews on cancer databases
In complement, and not to duplicate good recent publications in the last months, some reviews on cancer databases list most of the Internet resources in the general field of cancer genomics. A review of L. Chin gives an overview of the current state if cancer genomics (L. Chin et al., 2011). Regardless of a wide spectrum of references, the topic of cytogenetic resources is absent (Pavlopoulou A et al., 2015 ; Klonowska K et al., 2016 ; Brookes AJ, Robinson PN and Brookes AJ, 2015 ; Yang Y et al., 2015 ; Niroula A and Vihinen M, 2016 ; Diehl AG and Boyle AP, 2016; Martincorena I, et al. 2015). There are also many descriptions of database (and particularly in cancer) in all special issues of Nucleic Acid Reseach (each year in January).

2. GENERAL RESOURCES
Note: a detailed description of General resources in Genetics and/or Oncology may be found at href=http://atlasgeneticsoncology.org/Deep/General_ResoucesID20144.html

1. Bibliography
PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) is a widely used and free search engine and database of...
biomedical citations and abstracts, based essentially on the MEDLINE database of references on life sciences and biomedical topics. Medline is the U.S. National Library of Medicine (NLM) premier bibliographic database. PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/) is an archive of biomedical and life sciences journal literature. Articles are deposited by participating journals, as well as for author manuscripts that have been submitted in compliance with the public access policies of participating research funding agencies. Scopus (http://www.scopus.com/) is a database owned by Elsevier.

II- Nomenclatures
Gene Nomenclature: The HUGO Gene Nomenclature Committee (HGNC, http://www.genenames.org/) is the authority that assigns standardised nomenclature to human genes. Nomenclature for the description of sequence variations (http://www.hgvs.org/mutnomen/) is maintained by the Human Genome Variation Society (HGVS). International System for Human Cytogenetic Nomenclature (ISCN): The ISCN is the language used to describe abnormal karyotypes. International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3): The WHO/OMS has established a code, which provides a topographical (organ) identifier and an identifier for the detailed pathology.

III- Nucleic acid, genes and protein databases
Nucleic acid databases: GenBank (http://www.ncbi.nlm.nih.gov/genbank/) is a DNA sequence database. The need to have (in parallel to the genome projects) the best representation of genomic and transcript sequences (for diverse species) has been at the origin of consensus databases (as RefSeq, UCSC, Ensembl) with several methods of optimisation. Genomic sequences and transcripts: RefSeq (http://www.ncbi.nlm.nih.gov/refseq/) maintains and curates a database of annotated genomic, transcript, and protein sequence records. Ensembl (http://www.ensembl.org/) developed a software which produces and maintains automatic annotation on selected eukaryotic genomes. The UCSC Genome Browser database (see above) is a large collection containing genome assemblies of various species. Proteins: In addition to the amino acid sequence, protein name and description with domains, these databases may provide a brief annotation information, others are only computationally analysed. These databases are the following: UniProt (http://www.uniprot.org/), a hub consisting of two sections: "TrEMBL" and "Swiss-Prot"; neXtProt (http://www.nextprot.org/db/); PhosphoSitePlus (http://www.phosphosite.org/homeAction.action), an excellent resource providing comprehensive information and tools for the study of protein post-translational modifications; PROSITE (http://prosite.expasy.org/) Pfam (http://pfam.xfam.org/) and InterPro (http://www.ebi.ac.uk/interpro/). The Atlas of Genetics and Cytogenetics in Oncology and Haematology presents highly curated paragraphs with the description of the protein, but on a restricted sample.

IV- Cards
Entrez Gene (http://www.ncbi.nlm.nih.gov/gene/) is NCBI’s primary text search and retrieval system that integrates the PubMed database and molecular databases including DNA and protein sequence, structure, gene, genome, genetic variation and gene expression. Genecards (http://www.genecards.org/) is a database that provides information on all annotated and predicted human genes.

V- Genome cartography
The cartography of genes on a genome has always been a fundamental mean of representation of genomic information. With the human Genome Project, several types of viewers have been developed. To date, two sites are of first interest for human genetics: The UCSC Genome Browser website (http://genome.ucsc.edu/) contains the reference sequence for a large collection of genomes. The Genome Browser zooms and scrolls over chromosomes, "Blat" quickly maps a sequence to the genome. The UCSC Cancer Browser (https://genome-cancer.ucsc.edu/proj/site/haplot/)

VI- Structural variation databases
Genomic structural variation (including insertions, deletions, inversions, translocations and locus copy number changes) accounts for individual differences at the DNA sequence level in humans and can play a major role in diseases. Several databases have integrated data produced in the literature on copy number variation of DNA sequences: dbVar (http://www.ncbi.nlm.nih.gov/dbvar/), DGV - Genomic Variants (http://dgv.tcag.ca/dgv/app/home), DECIPHER (https://decipher.sanger.ac.uk/) and 1000 Genomes (http://www.1000genomes.org/).

VII- Polymorphism databases
It is important to distinguish polymorphisms due to single nucleotide (SNP) as the variability within a population and mutations acquired in a neoplastic process. The determination of variants was previously obtained by SNP arrays, but is nowadays performed by massive parallel sequencing. Polymorphism databases are: dbSNP (http://www.ncbi.nlm.nih.gov/SNP/overview.html), HAPMAP
Internet databases and resources for cytogenetics and cytophysiology


VIII- Portals/Working consortia

The primary goals of these projects are to generate catalogues of genomic abnormalities (somatic mutations, SNP genotyping, copy number variation profiling, abnormal expression of genes, epigenetic modifications) of series of genes in tumors from different cancer types. The main portals are: TCGA (http://cancergenome.nih.gov/), ICGC: (https://icgc.org/), OASIS (http://www.oasis-genomics.org/) and Firebrowse (http://firebrowse.org/).

IX- Pathology

Authoritative books in pathology includes clinical features, morphologic, immunohistochemical and molecular genetic features and prognosis, with a very large iconography. They are the following: the "Rosai and Ackerman's Surgical Pathology" and the "WHO/IARC Classification of Tumours series" (http://publications.iarc.fr/Book-And-Report-Series/Who-Iarc-Classification-Of-Tumours). The Armed Forces Institute of Pathology (AFIP) publishes series of the "AFIP Atlas of Tumor Pathology". The Atlas of Genetics and Cytogenetics in Oncology and Haematology provides complete description of diseases, but again on a limited sample; on the other hand, articles on genes closely related to these diseases are found, right next, in the Atlas. As a product of collaborative work, the usefulness of the Atlas is dependent on colleague participation in updating and completing it. PathologyOutlines (http://pathologyoutlines.com/) provides iconography. To be also noted, the United States and Canadian Academy of Pathology (USCAP, http://www.uscap.org/). The International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3) gives ICD-O codes for each cancer, with an ICD-O3-TOPO, which provides a topographical (organ) identifier and an ICD-O3-MORPH, which provides the basic and detailed pathology.

XI- Cancer Registries


XII- Patient associations and interfaces between science and patients - freely accessible services

Associations of patients and friends of patients: These associations of parents of patients with a rare disease are precious. They give moral support and help, and offer practical guidances and information about social benefits, subsidies and day-to-day life to families affected by illness. They often establish a program of grants for research (e.g. Xeroderma Pigmentosum Society (http://www.xps.org/), International Cancer Control (UICC, http://www.iacr.com.fr/), guidelines for registry practices and standard definitions, quality control, consistency checks and basic analysis of data, making data comparable between registries. The European Network of Cancer Registries (ENCN, http://www.encr.eu/) has the same role in Europe as IACR has worldwide. The National Program of Cancer Registries (NPC, http://www.cdc.gov/cancer/), maintained by the Centers for disease control and prevention (CDC), collects data on cancer occurrence in the USA. The Surveillance, Epidemiology, and End Results (SEER, http://seer.cancer.gov/) is a program of the National Cancer Institute. To be cited as well, the Union for International Cancer Control (UICC, http://www.uicc.org/).

3. CYTOGENOMICS RESOURCES

Note: a detailed description of Cancer Cytogenomics resources may be found at Cancer Cytogenomics resources

I- Chromosome rearrangements/Hybrid genes

Mitelman Database:

The database of chromosome aberrations in cancer counts a total number of cases amounting to more than 60,000, implicating more than 10,000 gene fusions, culled from the literature and organized into distinct sub-databases: The "Cases Quick Searcher" and the "Cases Full Searcher" contain the data related to chromosomal aberrations in individual cases. The "Molecular Biology Associations Searcher" collects cases according to the gene
rearrangements. The "Clinical Associations Searcher" is based on tumor characteristics, related to chromosomal aberrations and/or gene rearrangements. This free access database shows raw data and is reliable.

**Atlas of Genetics and Cytogenetics in Oncology and Haematology:**
The Atlas (http://atlasgeneticsoncology.org) is a peer reviewed on-line journal encyclopaedia and database with free access on the Internet. It is an integrated structure and comprises the following topics: genes, cytogenetics and clinical entities in cancer, and cancer-prone diseases. The Atlas combines various types of knowledge all on one site: genes, gene rearrangements, cytogenetics, protein domains, function, cell biology, pathways. It also contains clinical genetics, including hereditary diseases which are cancer-prone conditions, and diseases, focusing on cancers, but also listing other medical conditions. The Atlas is mainly composed of structured review articles or "cards" (original monographs written by invited authors). The Atlas contributes to the cytogenetic diagnosis and may guide treatment decision making!

**COSMIC** (http://cancer.sanger.ac.uk/cosmic) is a catalog of somatic mutations in cancer. It includes all abnormalities, from single nucleotide variations to chromosome rearrangements / fusion genes.

**Other resources:**
chimerDB
http://biome.ewha.ac.kr:8080/FusionGene/ is a database of fusion genes with PubMed references and some information about the structure of chimeric genes. TICdb (http://www.unav.es/genetica/TICdb/) is a database of Translocation breakpoints In Cancer with the fusion sequences at the nucleotide level. ChiTARS (http://chitars.bioinfo.cnio.es/) is a database of chimeric transcripts. TCGA Fusion gene Data Portal (http://54.84.12.177/PanCanFusV2/) presents an analysis across tumor types of the TCGA program. Other resources are OMIM (http://www.omim.org/), Fusion cancer (http://donglab.ecnu.edu.cn/databases/FusionCancer/), "Cancer Cytogenetics: Chromosomal and Molecular Genetic Abberations of Tumor Cells" is a book authored by Sverre Heim and Felix Mitelman.

**II- Data for SKY and FISH**
Fluorescence in-situ hybridization (FISH) technique enables identification of chromosomal structures to be identified using specific probes. This significantly improves the localisation of breakpoints on chromosomes. FISH technique can also be used on non-dividing cells (interphase nuclei). The Cancer Chromosome Aberration Project (CCAP) has generated a set of BAC clones that have been mapped cytogenetically by FISH and physically by STSs to the human genome. The BAC data is integrated into various databases (http://cgap.nci.nih.gov/Chromosomes/CCAPBAC Clones), (http://mkweb.bcgsc.ca/bacarray/). All BAC can be located on the UCSC genome browser (http://genome.ucsc.edu). BAC from the fishClones file can be visualized on the chromosomal bands on the Atlas (http://atlasgeneticsoncology.org/Bands/). More recently, several commercial companies have developed more specific catalogs of FISH clones as oligonucleotides probes.

**III- Comparative genomic hybridization (CGH) resources**
This technique detects disequilibria between a disease sample and a normal sample. Several sites are repositories for these CGH/SNP profiles: GEO, http://www.ncbi.nlm.nih.gov/geo/, Array Express (http://www.ebi.ac.uk/arrayexpress/), Tumorscape (http://www.broadinstitute.org/tcga/home), MetaCGH (http://compbio.med.harvard.edu/metasnp/), CaSNP (http://cistrome.org/CaSNP/), Cell line project (http://cancer.sanger.ac.uk/cell_lines), Cancer Cell Line Encyclopedia (http://www.broadinstitute.org/ccle/home) and ArrayMap (http://www.arraymap.org)

**IV- Mutation databases**
The determination of variants was previously obtained by SNP arrays, but is nowadays performed by massive parallel sequencing. As a result, a huge quantity of polymorphisms and mutations in tumors, are compared to controls. The landscape of the majority of recurrent mutations is now known and can be used for diagnosis. Even in haematological malignancies, where the chromosome rearrangements have shown to bear a major role, nonetheless, it appears now that some mutations at the nucleotide level can still be very important in determining treatments in relation to patient outcome (e.g. ASXL1, ATM, BCL6, BRAF, KRAS and NRAS, CBL, CCND3, CDKN2A and CDKN2C, CEBPA, CRLF2, ETV6, FLT3, GATA2, ID3, IDH1, IDH2, IKZF1, JAK1, KIT, MYD88, NOTCH1, NPM1, RUNX1, TP53). The main mutation databases are: COSMIC (http://cancer.sanger.ac.uk/cosmic), CENSUS (http://cancer.sanger.ac.uk/census/), HGMD (http://www.hgmd.cf.ac.uk/ac/index.php), LOVD (http://www.lovd.nl/3.0/home), TCGA eBioPortal (http://www.ebiportal.org/), IGCG Data Portal (https://dcc.icgc.org/), OASIS Portal (see above), IntOGen (http://www.intogen.org), BioMuta v2 (https://hive.biochemistry.gwu.edu/tools/biomuta/), DoCM (http://docm.genome.wustl.edu/), CIViC (https://civic.genome.wustl.edu/#/home), and ExAC (http://exac.broadinstitute.org).
<table>
<thead>
<tr>
<th>Resource</th>
<th>URL</th>
<th>Location</th>
<th>PMID</th>
<th>Last Update</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Human Gene Nomenclature Database (HGNC)</td>
<td><a href="http://www.genenames.org/">http://www.genenames.org/</a></td>
<td>(Huntsville, AL)</td>
<td>10017287</td>
<td>2005</td>
</tr>
<tr>
<td>Human Gene Nomenclature Society</td>
<td><a href="http://hgnc.biocury.com">http://hgnc.biocury.com</a></td>
<td>(Melbourne, Australia)</td>
<td>10295877</td>
<td>2005</td>
</tr>
<tr>
<td>Cancer portals</td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
</tr>
<tr>
<td>Genomic and Cytogenetic Resources</td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
</tr>
<tr>
<td>Cytogenetic databases and resources</td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
</tr>
<tr>
<td>Internet databases and resources for cytogenetics and cytogenomics</td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
<td><a href="https://www.genome.gov/39014683">List</a></td>
</tr>
</tbody>
</table>
4. PRACTICAL EXERCISE
See online Practical Exercises.

5. DISCUSSION
We have briefly discussed the various databases useful for clinicians, students, and researchers in finding answers to their questions and in determining which field in cancer research still needs to be studied. Only a handful of databases or portals take the cytogenetic information into consideration although being one of the first observation points confirming that the cell has transformed into a cancerous cell. Over the years (1960-2016), chimeric genes and fusion proteins have been discovered mainly by cytogenetic means. This has led to understanding of major cancerogenetic processes, and, later on, to the concept of treatments targets for many cases. Cytogenetics, or rather, cytogenomics of cancer, is a major contributor for the concept of “personalized medicine for cancer”, together with other tools for other mechanisms.

Dorothy Warburton wrote her colleagues in the American Cytogenetics Forum List: Subject: Is Cytogenetics Dying? “I have been told by my department chairman and other advisers that cytogenetics is dying at least three times: in 1968 (just before banding), in 1984 (just before FISH) and in 2001 (just before microarrays) (...) Chromosome changes are being recognised as the cause of more and more abnormalities, not fewer. This is true for both cancer and constitutional cytogenetics (...) There are also a great many basic facts about how chromosome abnormalities originate that we know little about and that are great fields for research”.

Besides what is known, there is still a lot to learn about how a mutation or a chromosomal rearrangement is influencing the cellular mechanism.

A decade ago, conventional wisdom was that the expressed genes were regulating all the processes in the cell and the rest of the genome was considered as “junk DNA”. The main purpose of this junk DNA was to act as a buffer and to protect the coding part from any kind of aggression. Progressively, we realized that the mutations happening outside the coding regions could influence the expression of nearby genes or in some cases several kilobases away. These mutations were located in regulatory regions recruiting or blocking transcription factors binding DNA.

More elements were found in the “junk DNA” with the discovery of the microRNAs (miRNA) and long non-coding RNAs (lncRNA). Both proved influential in regulating the genome and the expression of several genes. These miRNA and lncRNA are deregulated during the carcinogenic process and studies have demonstrated that by re-establishing the normal expression of one or few miRNAs or lncRNAs, the cancer cells transform back into normal cells. Afterwards, it was observed that genes which were not mutated could be silenced and vice versa. This was due to the epigenetic regulation of the genes and the recruitment of epigenetic regulators in different regions of the genome.

The cancerous cells actively modify their own epigenetic marks on the genome to increase the
expression of stemness genes providing them with proliferative or resistance advantages. The chromosome structure in the nucleus is dependent on many variables modulating the chromatin interactions across the whole genome. Techniques like chromatin conformation capture and interphase fluorescent in situ hybridization (FISH) detect spatial associations between specific genes. In the future, they are likely to be the techniques for detecting abnormalities associated with progression of tumours. Similarly, when modifying the 3D structure of chromosomes, it has been demonstrated that the disruption of chromosome neighbourhoods via mutations in insulated neighbourhood boundaries (cohesin CTCF interactions) activates proto-oncogenes in cancer cells (Hnisz D et al., 2016).

As mentioned above, the last build of the human genome (GRCh38, Dec 2013) is more precise than the previous ones and takes into account more haplotypes shared by parts of the human population. Nevertheless, due to a great part of repeated sequences of various classes (50% of the genome), there are some gene families only present at low frequencies, e.g. the insertions of endogenous retroviruses, which represent 8% of the genome and are mostly not localized on GRCh38 for a great part (Wildschutte JH et al., 2016). We need to recognize that such a diversity may be a factor leading to susceptibility to cancer or other diseases.

There is an increasing realisation that our environment, our nutrition and our way of life have an impact on how our cells mutate and how these mutations are repaired or how the abnormal cells are eliminated by our organism. It is difficult to fully understand and appreciate the complexity of the cellular mechanism and the various levels that can be deregulated in cancer makes. The use of databases makes this process easier as it condenses the complex information and provides links to other relevant databases providing even more specialized information.

Sequencing of thousands of genomes of patients bearing different types of tumours and genomes of normal persons, generates a huge amount of data demonstrating the complexity of tumours. Focus points are the sensitive parts of the genome which are more prone to be mutated, the expression profiles, the function of the proteins, the pathways the proteins are located in, and the protein interactions (in normal cases and abnormal cases). The same is done at the single cell level, bringing a huge amount of data demonstrating the heterogeneity of the tumour and the interaction with the niche. In principle, this wealth of integrated genomic data and clinical information could reveal the genetic bases of cancer, inherited diseases, and drug responses-illnesses and remedies that have touched nearly every person and family across the globe. This will constitute the databases of tomorrow. Interpreting these data requires a larger evidence base than any one party alone can develop. However, existing technologies, regulations and approaches are currently not designed for sharing and interpreting this wealth of information effectively, especially across diseases and nations. Databases will need to integrate information increasingly in the upcoming years but also stay interoperable with other databases. This corroborates the idea of having a common nomenclature and language to avoid mistakes but also have common and reliable information (http://www.aacr.org/Research/Research/Pages/aacr-project-genie.aspx#.Vx_mi8l9Qox; https://genomicsandhealth.org) in the same manner in which the Working Group for Planetary System Nomenclature maintains the “astronomical naming conventions and planetary nomenclature”, any scientific field, confronted with the exponential growth of data, has to create a system of classification and thesaurus for naming of new items, and nosological definitions. This often implies the use of a structured grammar and census of accepted terms/objects.

Resources such as the International System for Human Cytogenetic Nomenclature (ISCN), the International Classification of Diseases for Oncology (ICD-O), the Human Gene Nomenclature Database (HGNC), and the Nomenclature for the description of sequence variations (from the HGVS) are indispensable research tools allowing a common language.

Interoperability: In the context of interoperability between recent cancer projects, such as TCGA and ICGC as well as all data produced by thousands laboratories or hospitals, the Global Alliance for Genomics and Health (Global Alliance, https://genomicsandhealth.org/) (Lawler M et al., 2015) was created to accelerate the potential of genomic medicine. This association brings together over 375 leading institutions working together to generate a common framework of harmonized approaches in enabling the data sharing. In particular, it works to establish interoperable technical standards (standardised language and tools) to management of genomic and clinical data and for a better representation of genotype-phenotype associations.

Open data (open source, open hardware, open content, and open access) is a concept so dear to many, in particular in the medical and scientific world. This concept has recently had a renewed vigour with/since the advent of the Internet. Maintaining the data open and free remains a daily struggle. Everything has a cost, even a free database, from upkeeping and updating the database to
providing for the scientific staff necessary to produce expertized data. Data should remain freely available. However, a business model remains to be established. Although economic investment by the public sector would benefit the whole of mankind, as well as economically profitable in the end, most of the institutional stakeholders are gradually disengaging, and well-known databases are forced to beg for funds (see recent examples, useless to cite them!) or to disappear. This would be a regrettable drawback for the scientific and medical community - yet it may still happen.

References


Boveri T. Zur Frage der Entstehung maligner Tumoren 1914 Gustav Fischer

Nowell PC, Hungerford DA. A minute Chromosome in Human Chronic Granulocytic Leukemia Science 1960 132:1497

Caspersson T, Zech L, Modest EJ. Fluorescent labeling of chromosomal DNA: superiority of quinacrine mustard to quinacrine Science 1970 Nov;137(3599):762


Turc-Carel C, Philip I, Berger MP, Philip T, Lenoir G. [Chromosomal translocation (11; 22) in cell lines of Ewing’s sarcoma] C R Seances Acad Sci III 1983;296(23):1101-3


Sinclair PB, Nacheva EP, Leversha M, Telford N, Chang J, Reid A, Bench A, Champion K, Huntly B, Green AR. Large deletions at the t(9;22) breakpoint are common and may identify a poor prognosis subgroup of patients with chronic myeloid leukemia Blood 2000 Feb 1;95(3):738-43

Möller E, Hornick JL, Magnusson S, Veerla S, Domanski HA, Mertens F. FUS-CREB3L2/L1-positive sarcomas show a specific gene expression profile with upregulation of CD24 and FOXL1 Clin Cancer Res 2011 May 1;17(9):2646-56


Fletcher CD. The evolving classification of soft tissue tumours - an update based on the 2013 WHO classification Histopathology 2014 Jan;64(1):2-11


Hokland P, Ommen HB. Towards individualized follow-up in adult acute myeloid leukaemia in remission Blood 2011 Mar 13;117(9):2577-84


Diehl AG, Boyle AP. Deciphering ENCODE Trends Genet 2016 Apr;32(4):238-49


Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells Science 2015 Sep 25;350(6255):1483-9


Diehl AG, Boyle AP. Deciphering ENCODE Trends Genet 2016 Apr;32(4):238-49


Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells Science 2015 Sep 25;350(6255):1483-9


Mitelman F, Johansson B, Mertens F. Mitelman database of chromosome aberrations and genes fusions in Cancer


Amerberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders Nucleic Acids Res


Atlas Genet Cytogenet Oncol Haematol. 235
for precision medicine in oncology Ann Oncol 2015 Mar;26(3):582-8
Clough E, Barrett T. The Gene Expression Omnibus Database Methods Mol Biol 2016;1418:93-110
Cooper DN, Krawczak M. Human Gene Mutation Database Hum Genet 1996 Nov;98(5):629
Deng M, Brägelmann J, Schultzje L, Perner S, Web-TCGA: an online platform for integrated analysis of molecular cancer data sets BMC Bioinformatics. 2016 Feb 6;17:72