REVIEW

Human Mutation
OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Locus-Specific Databases in Cancer: What Future in a Post-Genomic Era? The TP53 LSDB paradigm

Thierry Soussi[1,2]*

[1] Department of Oncology-Pathology Cancer Center Karolinska (CCK), Karolinska Institute, Stockholm, Sweden; [2] Université Pierre et Marie Curie Paris 6, Paris, France

For the TP53 Special Issue

Received 13 November 2013; accepted revised manuscript 16 January 2014.
Published online 29 January 2014 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22518

ABSTRACT: Locus-specific databases (LSDBs) are curated compilations of sequence variants in genes associated with disease and have been invaluable tools for both basic and clinical research. These databases contain extensive information provided by the literature and benefit from manual curation by experts. Cancer genome sequencing projects have generated an explosion of data that are stored directly in centralized databases, thus possibly alleviating the need to develop independent LSDBs. A single cancer genome contains several thousand somatic mutations. However, only a handful of these mutations are truly oncogenic and identifying them remains a challenge. However, we can expect that this increase in data and the development of novel biocuration algorithms will ultimately result in more accurate curation and the release of stable sets of data. Using the evolution and content of the TP53 LSDB as a paradigm, it is possible to draw a model of gene mutation analysis covering initial descriptions, the accumulation and organization of knowledge in databases, and the use of this knowledge in clinical practice. It is also possible to make several assumptions on the future of LSDBs and how centralized databases could change the accessibility of data, with interfaces optimized for different types of users and adapted to the specificity of each region of the genome, coding or noncoding, associated with tumor development.

Hum Mutat 00:1–11, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: LSDB; TP53; p53; cancer; next-generation sequencing; biocuration

Introduction

Using protein sequencing, Vernon Ingram was the first to discover that a small change in the hemoglobin protein could lead to human sickle-cell anemia [Ingram, 1956]. Since this pioneering work, it has been largely demonstrated that gene mutations are the basis for most genetic diseases. (Although epigenetic modifications are also an important component in all human diseases including cancer, they will not be discussed here as they are not integrated in LSDBs.)In the late 1970s, the tedious task of protein sequencing was replaced by the revolutionary introduction of DNA sequencing and molecular cloning technologies [Collins, 1995]. Over the years, progress has been made in identifying the genes involved in both monogenic and polygenic disorders, including such complex diseases as cancer. For these genes, numerous and various types of alterations have been described, ranging from point mutations to large deletions or translocations. Reporting, storing, classifying, and analyzing these mutations have been a major challenge [Horaitis and Cotton, 2004]. For many years now, locus-specific databases (LSDBs) have been developed for this purpose. Although LSDBs are independently developed for single genes, they do offer great accuracy as they are curated manually by experts in the field [Claustres et al., 2002; Auerbach et al., 2011]. They provide information that can be used for large-scale analyses and often include structural, functional, or evolutionary data. For constitutional mutations associated with a genetic syndrome, several LSDBs also include phenotypic data useful for the study of genotype–phenotype correlation.

The development of high-throughput methodologies capable of analyzing the pattern of transcription of an entire cell or tissue (expression array or RNA sequencing), defining the structure and organization of the chromatin (chromatin immunoprecipitation) or sequencing an entire genome in a few days (next-generation sequencing, NGS) has radically changed the entire field of biology [Park, 2009; Hawkins et al., 2010; Metzker, 2010; Ozsolak and Milos, 2011]. Furthermore, this methodological revolution led to the discovery of novel layers of complexity in gene organization and how their expression is regulated. These spectacular advances have laid a path to a postgenomic era in which healthcare research and provision will be very different.

In cancer research, genomic studies in the pregenomic era were limited; they were focused either on a small number of genes analyzed in large patient cohorts, or on a more significant number of genes but in only a few tumors (Fig. 1). Indeed, large-scale analysis combining a multitude of genes and tumors was a Herculean and costly task.

Today, in the postgenomic era, these barriers have fallen and whole genome sequencing in a multitude of tumors can be performed in a matter of weeks. The International Cancer Genome Consortium (ICGC, http://dcc.icgc.org/), The Cancer Genome Atlas Project (TCGA, http://cancergenome.nih.gov/), and the Sanger Institute (http://www.sanger.ac.uk/) have undertaken large-scale cancer genome analyses in different types/subtypes of cancer and several reports from these projects have already been published [Hudson et al., 2010; Alexandrov et al., 2013, Garraway and Lander, 2013, Koboldt et al., 2013]. These studies will lead to profound changes in LSDB management.

*Correspondence to: Thierry Soussi, Department of Oncology-Pathology Cancer Center Karolinska (CCK), Karolinska Institute, Stockholm, Sweden. E-mail: thierry.soussi@ki.se

© 2014 WILEY PERIODICALS, INC.

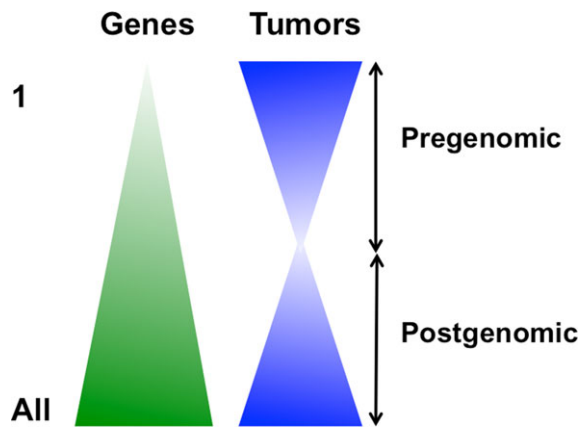

Figure 1. Evolution of sequencing strategies used in tumor analysis. The use of conventional Sanger sequencing restricted studies to a few genes in a large number of tumors; studies covering a multitude of genes were scarce. Large-scale analyses of multiple genes began by targeting specific gene families such as protein kinases. With the advent of NGS, whole genome or exome sequencing of a large number of tumors is now feasible at a reasonable cost.

Whole tumor genome sequencing will result in an enormous increase in the volume of acquired data and lead molecular genetics to a new world of "big data" [Schadt et al., 2010]. Exaoctets of raw data (1 Gigaoctet (Go) = 10^9 octets; 1 Exaoctet (Eo) = 10^{18} octets) will be generated and sophisticated software will be required for mining and interpreting them. Handling, accessing, and analyzing this amount of data will also require novel computing processes such as cloud or heterogeneous computing. Furthermore, the extraction of specific gene information will be a very challenging task for LSDB curators [Metzker, 2010, Schadt et al., 2010].

Recent studies have shown that the organization of the mammalian genome is far more complex than previously thought and that noncoding regions of the genome are also potential targets for alterations [Gerstein et al., 2012]. Evaluation of mutation pathogenicity must therefore evolve and take into consideration a setting larger than the end protein [Sauna and Kimchi-Sarfaty, 2011]. Furthermore, cancer genomes are polluted by thousands of random, "passenger" mutations unrelated to neoplastic progression, necessitating novel bioinformatics tools to identify the few relevant "driver" mutations truly associated with cell transformation [Chanock and Thomas, 2007].

How will the definition of a cancer mutation evolve in the future? What will be the fate of LSDBs in the postgenomic era? Will they even survive these major changes? How will the flow between data, curators, and users be modified by these new technologies, which displace publications as the primary material for data mining? To address these questions, the present review will focus on the *TP53* gene (MIM #191170). Indeed, this latter is the most frequently mutated gene in human cancer and the TP53 database is the largest collection of mutations, furthermore compiled from a wide gamut of cancer types.

Thus, the evolution and content of the TP53 LSDB will be used here as a paradigm to explore the current situation and make assumptions for the future of LSDBs. This review will focus

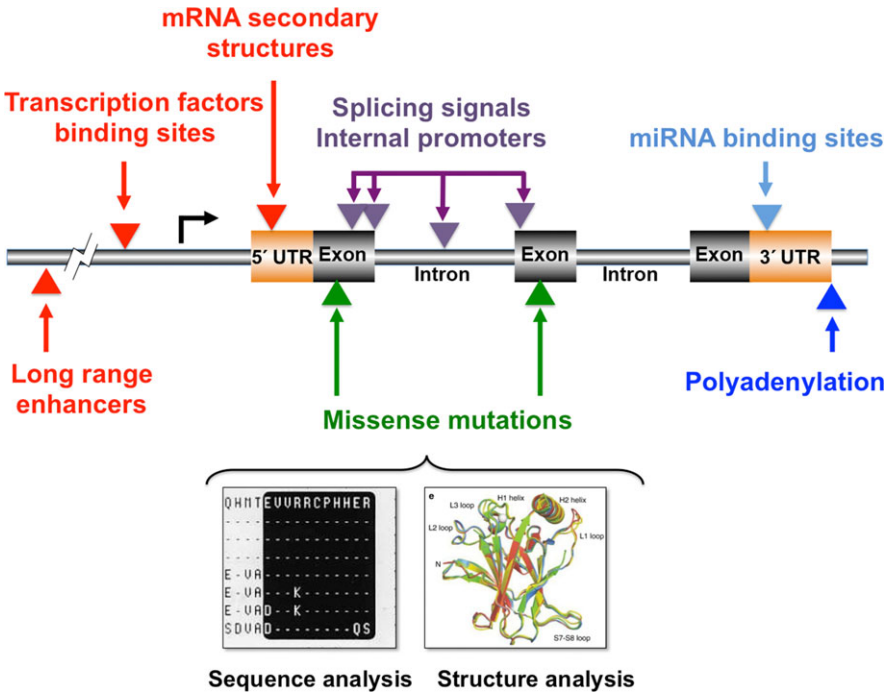

Figure 2. Consequences of functional alterations in various regions of the genome. Eukaryotic genes contain multiple regulatory sequences that may be selected by mutations in cancer. The 5′ end includes both distal and proximal regulatory sequences that recruit the transcriptional machinery. The 5′UTR can adopt multiple secondary structures that regulate protein translation. Splicing signals can be found in different regions of the introns, as well as in exons with ESE (Exonic Sequence Enhancer) or ESS (Exonic Sequence Silencer) sequences. The 3′UTR region contains multiple potential binding sites for various miRNA that regulate mRNA fate. The 3′ end region includes the polyadenylation sites and the transcription terminators. Several bioinformatics programs are available to detect potential deleterious mutations in all of these nontranslated regions. Mutations in coding regions are more easily validated via computational tools based on sequence or structural analysis. Figure adapted from Lee et al. (2009) with permission.

exclusively on cancer genes and cancer LSDBs as they have specificities and requirements generally different from those of other genetic diseases, although some issues are also applicable outside the strict setting of cancer. Recent works have demonstrated the presence and predominance of passenger mutations in tumor genomes. These passenger mutations must be differentiated from the rare driver mutations, a challenge very similar to that faced by clinical geneticists to differentiate SNPs from disease-causing mutations. Germline mutations in tumor suppressor genes lead to hereditary cancer syndromes with heterogeneous penetrance managed as complex Mendelian disorders.

Issues related to the storage and management of large amounts of data will not be discussed here as they have already been extensively reviewed [Schadt et al., 2010; Hood and Rowen, 2013].

Mutations in the Pre- and Postgenomic Eras

Cancer is a genetic disease characterized by a wide variety of genomic alterations, ranging from single-nucleotide substitutions to large chromosomal abnormalities such as translocations or duplications [Stratton, 2011; Vogelstein et al., 2013]. Large-scale analyses of several thousand different types of tumors have revealed that tumoral DNA is flooded with single-nucleotide variants (SNVs) scattered across the entire genome, including both the coding and noncoding regions. Changes to DNA methylation, a process involved in the regulation of gene expression, are also key players in cancer.

The classical view prevailing forty years ago divided the genome into two entities: one comprising genes and their regulatory regions responsible for encoding messenger RNA translated in turn into protein; and one comprising "junk DNA" with unknown and consequently nonessential function [Susumo, 1972]. The resulting gene-centric view created an analytical bias toward coding regions. Identification of an SNV as an acquired somatic mutation was checked either by looking at the constitutional DNA or, more often, by comparing the sequence to the dbSNP, which gathers all natural polymorphisms found in the human genome (http://www.ncbi.nlm.nih.gov/SNP/). SNV were dichotomized into synonymous SNV (sSNV) that do not change the protein sequence and nonsynonymous SNV (nsSNV) that lead to an amino-acid substitution. These nsSNVs are the most deleterious for gene function. Nevertheless, a significant number of sSNV detected in exonic sequence should not be considered as silent mutations as they can lead to aberrant RNA splicing or structure and thus a decrease in protein translation [Sauna and Kimchi-Sarfaty, 2011]. The sSNV c.375G>T (p.=T125T) in the TP53 gene had been initially considered as neutral in multiple reports but was later shown to be detrimental for TP53 splicing [Varley et al., 1998].

Recent discoveries have profoundly changed our knowledge of the mammalian transcriptome and genome plasticity and oblige us to reconsider the definition of "cancer associated mutation." This new knowledge will also have important consequences on mutation analysis and interpretation (Fig. 2).

Complexity of the Human Genome and Transcriptome

The recent advent of high-throughput RNA sequencing and the ENCODE project have uncovered new layers of gene expression regulation and highlighted the extreme complexity and versatility of the genome [Gerstein et al., 2012; Pennisi, 2012]. The majority of human genes encode multiple transcripts through the use of alternative promoters, alternative splicing, or alternative polyadenylation. The combinatorial mechanism of alternative splicing increases the coding potential of the genome by allowing the synthesis of multiple protein isoforms with different—even antagonistic—functions from a single gene. Therefore, an nsSNV, depending on its location, may affect either the entire pool or only a subset of isoforms, leading to a wide variety of phenotypes. Several cancer genes targeted by SNV such as *RET* (MIM #164761), *BCL11A* (MIM #606557), TP53, or *BRCA1* (MIM #113705) encode multiple protein isoforms. Only 80% of TP53 mutations target the 12 protein isoforms; indeed, mutations in exons 2–4 will spare the six proteins that are missing the amino-terminal domain (see Soussi et al. in the same issue).

Gene expression is tightly regulated at multiple levels. At transcription, the coordinated interactions between extra and intragenic *cis*-acting elements and their associated *trans*-acting factors regulate tissue-specific and temporal gene expression. *Cis*-acting elements such as core and proximal promoter elements are typically restricted to within a couple hundred base pairs from transcriptional start sites and regulate genes in their immediate vicinity. In contrast, distal *cis*-elements are usually located at >1 kb and in some cases up to 1 Mb in either direction from a transcription start site. Additionally, there is evidence to suggest that genes can also be regulated in *trans* by elements on other chromosomes. However, because of the coding region bias mentioned earlier, analyses of cancer mutations often disregarded all these regulatory regions. Although knowledge of telomerase hyperactivity in human cancer has existed for two decades, it is only in recent studies that activating mutations in the proximal promoter of the *TERT* gene (MIM #187270) have been discovered [Horn et al., 2013; Huang et al., 2013]. The estrogen receptor, tightly associated with breast cancer, mediates its effects via distal elements that may be located 100 kb upstream of the transcription start site. Many genes regulated by the transcription factor TP53, such as *MDM2* (MIM #168745) or AIP1 (MIM #605426), contain a response element localized in intronic sequences. Germline SNPs have been discovered in a few of them and may lead to a heterogeneous TP53 response and some marked susceptibility to cancer [Bond et al., 2004; Tomso et al., 2005; Zeron-Medina et al., 2013]. Somatic mutations have not been reported yet, but lack extensive analysis due to their large number, high degenerescence, short size, and heterogeneous location in noncoding regions.

More recent transcriptomic research has brought to light a novel class of nonprotein coding transcript (non coding RNA, ncRNA) encoded by intergenic sequences previously defined as junk. The number of identified genes encoding ncRNA has increased exponentially (more than 80% of the human genome) over the past decade and now far exceeds that of protein-coding genes (less than 3% of the human genome) [Bernstein et al., 2012]. These ncRNA have multiple functions in the regulation of the transcriptional networks of mammalian cells. Dysregulation and alterations of genes encoding ncRNA have been identified in various types of diseases including cancer, but it remains to be known whether or not they outnumber mutations in coding regions. Among the various ncRNA, microRNA (miRNA) in particular have been extensively analyzed [Yates et al., 2013]. Their activities rely on hybridization to target sequences, called miRNA response elements (MRE), usually located on the 3′ end of mRNA. This binding leads to a reduction of protein synthesis via multiple mechanisms such as reduced translation or mRNA degradation. As there is no requirement for perfect identity with the target, miRNA can regulate the fate of multiple transcripts that are difficult to identify. Although miRNA are small molecules (22 nucleotides), they are derived from larger precursors via a complex maturation process. Therefore, SNV can be potentially localized in any sequence that will affect quantitatively or qualitatively miRNA activity. Furthermore, because of its mechanism of action, both the

miRNA and its potential targets can sustain pathogenic SNV. Thus, mutations in the 3′ UTR of many genes, previously discarded as passenger mutations, must be reassessed. A recent study on B-cell lymphoma uncovered somatic SNV in the 3′UTR of the *TP53* gene in more than 50% of patients [Li et al., 2013]. These mutations were localized in potential MRE for various miRNA that regulate TP53 mRNA fate.

Plasticity of the Human Genome

Somatic mosaicism denotes the presence of multiple populations of cells with different genotypes in a single organism [Biesecker and Spinner, 2013]. It was long assumed that somatic mosaicism was rare, occurring only occasionally in hereditary diseases (e.g., forms of Turner's syndrome, trisomy). Recent studies have revealed that mosaicism is far more frequent than previously thought and could concern the majority of somatic cells [Stratton et al., 2009; Jacobs et al., 2012]. Starting at fertilization and throughout life, the genome is continuously the target of exogenous or endogenous mutagens, DNA replication errors and various types of recombination (Fig. 3). The pattern and the frequency of these alterations differ according to cell type and to the degree of exposure to various exogenous mutagens, for example, lung epithelial cells exposed to tobacco smoke, skin cells exposed to UV light, or organ tissues exposed to gamma radiation. The vast majority of these somatic alterations are thought to be neutral with no phenotypic contribution. However, they do contribute to the constitution of a heterogeneous somatic mosaicism (Fig. 3).

In human cancer, when a clonal alteration with a selective growth advantage appears, a population with a specific genotype, differing slightly from the primordial genotype, will emerge (Fig. 3). Furthermore, during the transforming process, the mutation rate will continue with the same frequency or even accelerate if repair systems are impaired. In the pregenomic era, this heterogeneity could escape detection by being located in regions of the genome not screened for mutation and/or present in only a few percent of tumors cells, thus not picked up by global genome analyses (Fig. 3). Today, the ability to detect tumor heterogeneity and reconstruct the evolution of the various molecular events in single tumors via NGS and bioinformatics has shed new light on the importance of passenger mutations. However, we currently have no knowledge as to how many of these mutations reflect in reality somatic mosaicism before transformation [Campbell et al., 2008; Yates and Campbell, 2012]. Despite their apparent lack of clinical value, these mutations deserve attention because they exhibit specific mutational signatures highly related to the type of cancer, which is useful information for tracking cancer mutation etiology [Alexandrov et al., 2013, Kandoth et al., 2013, Lawrence et al., 2013]. Furthermore, it remains possible that some passenger mutations randomly selected in primary tumors will come into play later, for example, in the development of resistance to cancer treatments. From a semantic point of view, these mutations are neither driving mutations as they do not participate in transformation, nor passenger mutations as the may be ticking bombs waiting for specific events to occur [McFarland et al., 2013].

Passenger mutations can be found in coding and noncoding regions of the genome. They may also be difficult to distinguish from driving mutations, although this distinction is essential for obtaining an accurate picture of the cancer genome. For coding regions, several statistical approaches have been developed to solve this problem, for example, the comparison of observed and expected ratios of synonymous versus nonsynonymous variants. Alternatively, various bioinformatics methods can be used to indicate whether an amino-acid substitution is likely to damage protein function on the basis of either conservation through species or whether or not the amino-acid change is conservative in combination with other information including structural information obtained from the 3D-structure/homology models [Grantham, 1974; Adzhubei et al., 2010; Ng and Henikoff, 2006; Davydov et al., 2010]. The classification and pathological assessment of SNV in ncRNA genes will be far more complex than it is in coding regions and require better knowledge of ncRNA function and an accurate evaluation of the impact of mutations affecting them [Cooper and Shendure, 2011]. Novel sophisticated bioinformatics tools and carefully curated databases will be necessary to achieve this goal [Khurana et al., 2013; Lawrence et al., 2013]. The complexity of the mutational landscape and the expansion of selected targets to large, unexplored regions of the genome will be important challenges in cancer genetics. The NIH have recently launched several application calls for research projects focused on developing computational approaches for interpreting sequence variants found in the nonprotein-coding regions of the human genome (http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-13--013.html)

Discovery and Publication of Cancer Gene Mutations: The TP53 Paradigm

Upon the discovery of a novel cancer gene, a unique three-phase pattern is observable in publication, that is, a discovery phase, a consolidation phase, and an application phase. The length of the phases, individually and collectively, depends on the popularity of the gene, the type of alteration and its clinical relevance (Fig. 4A and B). During the discovery phase, the publications describe precisely novel mutations and discuss their potential pathogenicity in relation to the disease. A burst of studies then leads to the identification of novel mutants and their diversity rises quickly (Fig. 4A and B). This phase is commonly associated with reports published in journals with a high impact factor and parallels the rate of mutant or clinical novelties. Transition to the consolidation phase occurs quickly when genetic and clinical data become redundant. During this phase, the number of reported new mutants will decrease and the sequencing of multiple new clinical specimens will discover mostly previously described mutants, leading to a plateau in mutant novelty (Fig. 4A and B). This consolidation phase is vital as it adds nuance to and validates data from the discovery phase in a wide variety of clinical or geographical settings. Consequently, mutations are either described in supplementary materials or quoted as unpublished data, leading to a decrease in reported mutations. Except for a few very specific cases, the consolidation phase is accompanied by a decrease in the impact factor of the publishing journals. This decrease in descriptions of mutations does not reflect their frequency in the disease or the incidence of their analysis but rather a lack of interest and/or utility in their publication. If the mutation has no clinical value, the number of studies will drop quickly then stop. It has also been observed that the consolidation phase is associated with an increase in inconsistent studies. An extensive analysis of the various flaws associated with the publication of mutations is provided by Kern and Winter in their 2006 review [Kern and Winter, 2006].

Finally, for (Fig. 4A) mutations with clinical value, a long application phase then begins. However, publications fall off as service laboratories do not consider reporting to be an essential part of their work and descriptions of novel mutations become scarce.

This model is well illustrated by an analysis of the *TP53* gene mutation database and the frequency of publications reporting TP53

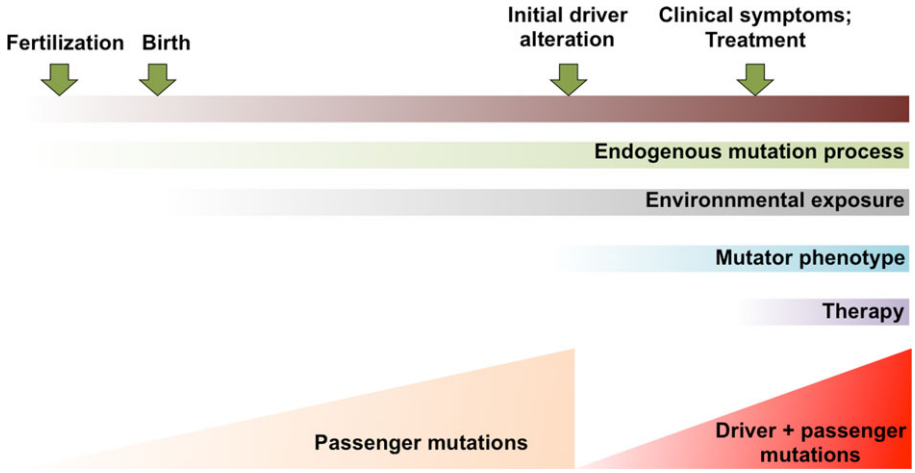

Figure 3. Plasticity of the human genome from fertilization to neoplasia. As early as fertilization, the human genome is subjected to a range of modifications, including copy number variations, recombinations, and single-nucleotide substitutions. Following birth, the lifestyle of the individual will have a profound impact on the frequency and pattern of exposure-dependent genome mutations. As long as these mutations are situated in genomic regions that do not induce a change in cell fitness, they will remain scattered randomly in multiple cells of the organism. In tumorigenesis, an initial driver mutation will occur in a cell with a genome differing from that of the zygote and containing a specific set of passenger mutations that will be coselected throughout transformation. Over time, new passenger and driver mutations will continue to emerge. At diagnosis, the tumor genome will have become highly heterogeneous, with a handful of driver mutations flooded in hundreds of thousands of passenger mutations. Defects in the various DNA repair pathways will also contribute to the increased frequency of mutations during transformation. Although passenger mutations occurring before the first driver mutations and the first round of selection will be detected in most tumor cells, later ones will be present only in subclonal populations. Upon treatment, novel mutations induced by radiation or chemotherapy may increase the genetic diversity of the tumors. In this simple model, passenger mutations are thought to be simple hitchhiking events without any role in transformation. Nevertheless, it remains possible that some passenger mutations arising before the first driver event will become advantageously or disadvantageously active at a later stage, leading to a complex pattern of mutations.

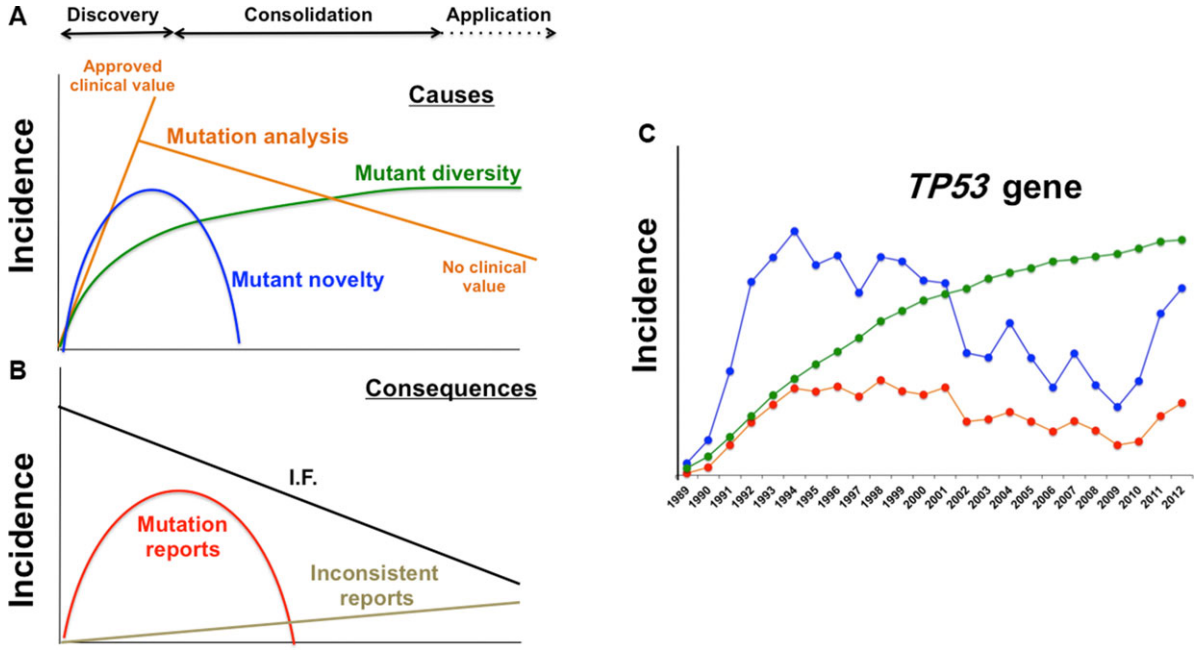

Figure 4. Evolution of gene mutation publications. **A and B:** During the discovery phase, the number of studies increases rapidly (orange line) leading for the most part to the identification of novel mutants (blue line) and mutant diversity rises quickly (green line). This is associated with a high frequency of published reports (red line) in high-ranking journals (black line). During the consolidation phase, mutant diversity reaches a plateau (green line), the number of mutation reports drops with a decrease in publication quality (black line) and an increase in inconsistent reports (brown lines). The application phase will begin only if the gene has clinical value. TP53 mutations are a good example of this model, as shown in panel **C**. These mutations began being published in scientific journals in 1989. The number of reports then accelerated quickly, reaching an apogee between 1997 and 2000 (red line). This paralleled the description of the majority of mutant TP53 (blue lines). In 2000, 68% of the total number of mutations included in the TP53 database had already been identified (blue line). This number reached 80% for missense mutations as the majority of new published mutations were frameshift mutations, which are more metamorphic than missense mutations. Since 2011, there has been a slight rebound in mutation reports due to the sequencing of more than 3,000 tumor genomes via NGS.

mutations (Fig. 4C). The discovery phase began in 1989 with the first description of TP53 mutations in lung and colorectal cancer [Baker et al., 1989; Takahashi et al., 1989]. Over the following years, there was a constant increase in publications describing novel TP53 alterations in various types of cancer (http://p53.fr). Thus, culminated in 2001 with more than 2,000 mutations reported in 300 publications [Soussi and Beroud, 2001; Soussi et al., 2006]. More than 85% of the mutant TP53 listed in the database was identified during the discovery phase. The decline in published TP53 mutations began in 2000, corresponding to the beginning of the second, consolidation phase. This decline was because of the difficulty of publishing TP53 mutations in peer-reviewed journals as their novelty wore off. In recent publications, TP53 mutations are not described because of journal space considerations. This trend toward the nonreporting of mutations is not specific to TP53; it applies to most cancer genes and raises important issues for the evolution of cancer LSDBs as they rely predominantly on published materials.

Whole tumor genome sequencing is currently accelerating the discovery of new cancer genes. Since the pioneering discovery of BRAF (MIM #164757) mutations in melanoma by Davies et al. (2002), numerous genes have been shown to carry mutations in various types of cancer [Davies et al., 2002]. For many of these genes, the discovery phase is just getting underway and several questions, such as (1) the frequency of the mutations, (2) the various types of cancer targeted by the alteration, and (3) their clinical relevance, are still awaiting answers.

These studies shed light on genes and pathways that were not previously under extensive analysis. This is best exemplified by the discovery of isocitrate dehydrogenase 1 and 2 (*IDH1*, MIM #14700 and *IDH2*, MIM #147650) mutations in gliomas [Parsons et al., 2008]. These enzymes are part of a multienzymatic complex localized in either the cytoplasm (IDH1) or the mitochondria (IDH2) and participate in the Krebs cycle, an essential metabolic pathway. IDH1 mutations are rare in primary glioblastoma multiforme but common in secondary glioblastoma multiforme indicating that they could be useful markers for glioblastoma stratification. A better prognosis for patients with IDH1 mutations has been observed in several studies but its clinical utility has not yet been established [Gupta et al., 2011]. The frequency of IDH1 mutation in other types of cancer must be confirmed but novel IDH1 mutants have not been identified as the majority of mutational events are restricted to a few codons. Analysis of IDH1 and IDH2 is now in the consolidation phase where the utility of IDH mutations as clinical biomarkers will be studied. Furthermore, other genes are currently at the beginning of the discovery phase. For example, the recent discovery of *TRAPP* and *GRIN2* mutations in melanoma or *GATA3* (MIM #131320) in breast cancer will open several new fields of investigation as these genes were not previously associated with cancer [Wei et al., 2011; Banerji et al., 2012].

The pace of the discovery phase accelerated rapidly with the release of the sequences of several thousand cancer genomes. These sequence analyses not only confirmed the participation of the usual culprits, such as *KRAS* (MIM #190070), *PIKC3A* (MIM #171834), or *TP53*, but also led to the discovery of a multitude of new suspects, with furthermore enough evidence to identify some of them as true driver genes. The genes *ARID1A* (associated with chromatin modeling, MIM #603024) and *GATA3* (associated with differentiation) are among the few to have been validated to date; others, with lower frequencies of mutation, will need more research.

As discussed by Wood et al. (2007), genes with high frequencies of mutation (gene "mountains" according to those authors) should be easy to identify but will be more so the exception rather than the rule. Inversely, genes with low frequencies of mutation (gene "hills") will be far more difficult to distinguish from neutral passenger mutations. Furthermore, because of the stochastic nature of mutations, the multiplicity of the various pathways, and the enormous cross-talk between them, we can anticipate that some cancer genes will be mutated very infrequently. As recurrence is a strong criterion for inferring the relevance of a mutation in cancer, a novel mission, named "the 10K project," is being launched to sequence 10,000 tumors per cancer type with the goal of uncovering very rare cancer genes with sufficient statistical power (http://news.sciencemag.org/biology/2013/03/ready-more-10000-cancer-genomes-projects). With the forthcoming release of the next-generation sequencer, the increasing power and accuracy of computational tools and the completion of the various large-scale cancer genome projects, it is not unrealistic to predict that the majority of cancer genes targeted by small alterations will be identified within the next decade. As these works progress, the accessibility of multiple cancer genome sequences will increase tremendously, and the discovery phase for any newly detected cancer genes will be shortened as their status in other cancers will be quickly determined by data mining. Ultimately, after completion of these sequencing projects, we will reach a plateau where the majority of gene regions involved in cancer will be identified, including a catalog of the SNVs associated with each one. Thus, a full molecular portrait of various types of human cancer will be available. Consequently, medical research will enter a new era where the emphasis is placed on clinical studies and finding the relevance of these data to improve patient care.

The Future of Cancer LSDBs

The explosion of information described in the previous sections will have profound consequences on LSDBs. To date, the majority of cancer gene LSDBs are simple lists of mutations that are very difficult to search. They are also highly heterogeneous in terms of quality, content, and format. Several database management systems, such as the Universal Mutation Database, the Leiden Open Variation Database, and the MUTbase have been developed to standardize the current data via a framework for LSDB curation but these systems are not intercompatible [Auerbach et al., 2011]. The human Genome Variation Society has published several guidelines on database structure and content or on variant nomenclature but with little in-the-field success [Kohonen-Corish et al., 2010]. Several recent surveys noted that less than half of the current LSDBs would meet minimal criteria for ease of use. Furthermore, the international nomenclature for publishing DNA variants (http://www.hgvs.org/mutnomen/recs-DNA.html) is used in less than 20% of publications, leading to data that are often useless for inclusion in LSDBs (Soussi, unpublished observations).

Most of the present LSDBs followed a similar development pathway. The data included in them were generally derived from publications, personal results from the curators or unpublished results from a consortium of specialists (Fig. 5).

LSDBs are maintained by one or several curators who are specialists in the field and thus benefit from strong scientific expertise. Data mining for LSDBs was historically performed manually and was task intensive. To ease the process, several procedures were established. First, algorithms were developed to identify publications that describe variants associated with a specific gene name. Second, tools for extracting mutation data from publications automatically were developed to circumvent error-prone manual entry of mutations in the database. However, these procedures function only with articles that respect the international mutation nomenclature and furthermore they do not distinguish tables that describe mutations

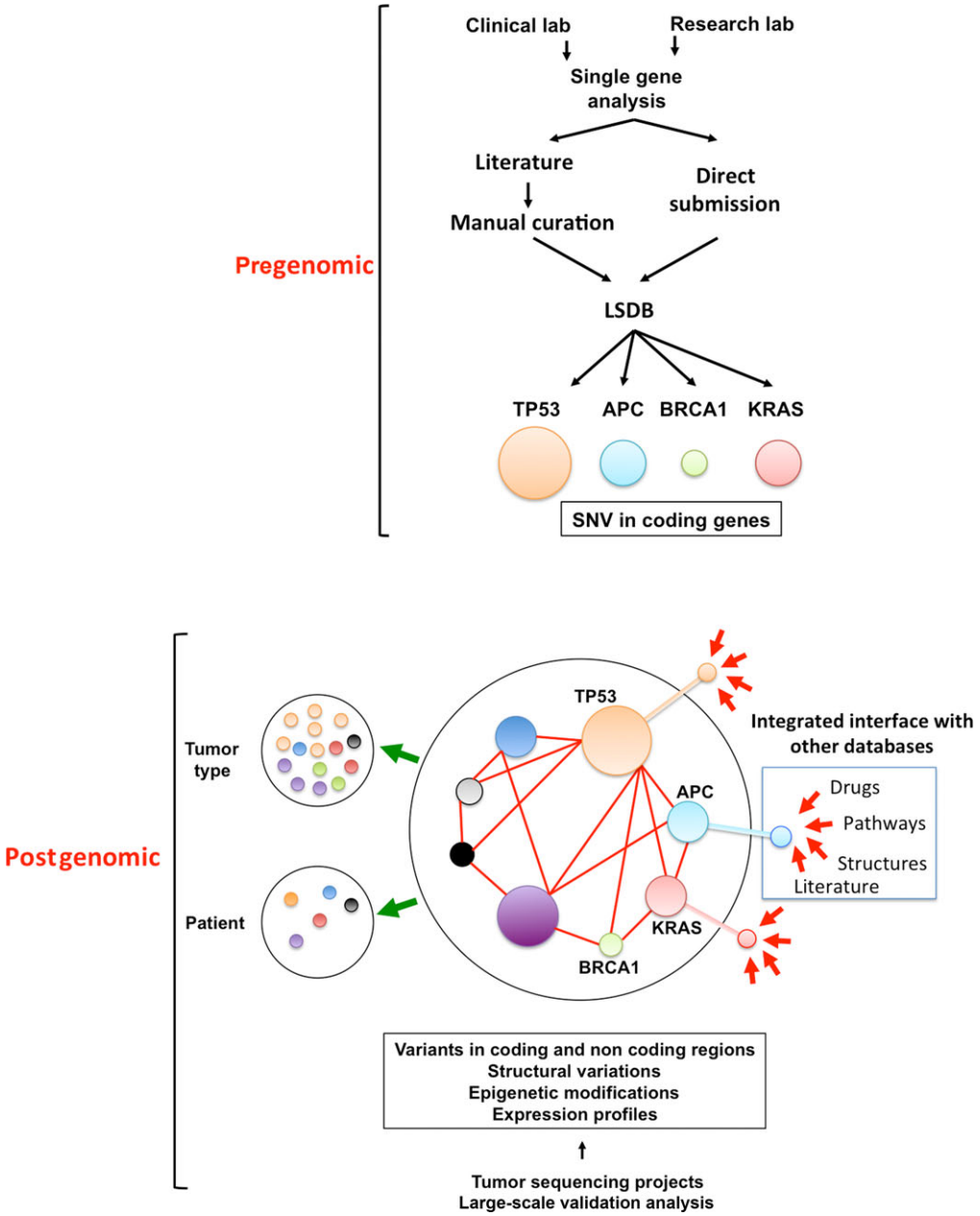

Figure 5. Cancer gene databases in the pre and postgenomic eras. Before the flood of data provided by the tumor sequencing projects, cancer gene LSDBs were built from data obtained from the literature. They were highly heterogeneous in size, format, and maintenance. Integrated analyses of multiple genes were not possible. The new centralized cancer gene databases gather data from tumor sequencing projects or large-scale validation analyses. These new databases will permit the identification of the most significant alterations for various types of cancer (tumor type analysis) or the identification of specific pathways (patient analysis). See text for more information.

in more than one gene. Maintenance is another problem for these LSDBs: more than 50% of LSDBs received regular maintenance and updates for only a few years after their launch. Indeed, maintenance of single LSDBs is time consuming, not rewarding, and poorly funded.

With the release of whole cancer genome sequences, this development pathway will face obsolescence as data are directly included in centralized databases such as those maintained by the Welcome Trust Sanger Institute (COSMIC database, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/), TCGA, http://cancergenome.nih.gov/), or the ICGC (http://dcc.icgc.org/). Although mutation data currently remain available and extractable from supplementary materials, this source will soon run dry, as all data will be mined from centralized databases in a near future. The aggregation of all cancer gene mutations in individual databases has multiple advantages, but how the data are organized in any one database will dictate the type of studies that can be realized based on it. The TCGA database provides a patient- and tumor type-centric organization that allows multigene analysis in individual patients. Furthermore, multipatient analyses will allow for the determination of statistically significant co-occurrence or co-exclusion of mutations situated in different genes, an important feature for

analyzing the specific pathways targeted by these mutations in a specific type or subtype of cancer. The ICGC and TCGA databases only include data obtained via the large sequencing programs of their parent consortiums (Fig. 5). The COSMIC database provides a gene-centric organization providing an accurate description of the various mutational events that target each cancer gene in different types of cancer. Its data are harvested from the literature and are not restricted to specific studies or projects. In contrast, this database is less opportune for tumor-oriented analysis and an absence of literature curation may permit the presence of spurious data.

We are experiencing a transition period where it is difficult for biocuration efforts to keep pace with mutation discovery, leading to large uncertainties in the information included in the various databases [Howe et al., 2008; Meyerson et al., 2010; Sanderson, 2011]. Overlap between the databases is not yet optimal and a single published study can lead to different sets of data in the three databases due to heterogeneous curation procedures. This can be illustrated by an analysis of mutations in several genes such as *TTN* (MIM #188840) and *MUC16* (MIM #606154), which encode the longest proteins of the human body. Mutations in *TTN* and *MUC16* have been reported in multiple tumor types and the COSMIC database reported 1,500 and 1,000 mutations for these two genes, respectively. The large size of these two genes and the lack of direct connections of their functions to cell transformation were highly suggestive of passenger mutations. The development of novel algorithms to infer the significance of cancer gene mutations has led to the discovery of multiple false-positive genes including TTN and MUC16 (Fig. 6) [Lawrence et al., 2013]. Mutational data also fluctuate depending on the algorithm used for variant calling. In several TCGA studies, the status of TP53 has been found to be heterogeneous depending on the database and the procedure used for the analysis (Soussi, unpublished observation, see also Leroy et al. in the same issue).

These uncertainties related to centralized databases, due to global automation and the absence of gene specific expertise, are far more important than those found in LSDBs. However, these issues are innate to the current, unstable transition period, which will surely pass notably as curation algorithms are developed to create accurate mutation databases. The MutSig algorithm of the Broad Institute provides a good example of how quickly algorithms can evolve (Fig. 6).

The most important question to be addressed is whether there is a future for independent cancer LSDBs. In the past, LSDBs have been tremendously useful. They benefit from rigorous expert curation, often coordinated by collaborating researchers with scientific expertise. Impaired gene functions caused by deleterious mutations and associated with specific phenotypes have generated working hypotheses and led to multiple lines of study. The four highly conserved domains of TP53 were identified as early as 1987 and the hot spot mutations within them in 1989. However, the DNA binding activity of this core (Soussi et al., 1987; Baker et al., 1989; Takahashi et al., 1989; Kern et al., 1991) region of the protein was not discovered until 1991.

Notch mutations are localized in various subdomains of the protein in different types of cancer, generating promising leads on the diversity of this pathway (Guruharsha et al., 2012). LSDBs have also been used for the development of diagnostic tools targeting the most significant mutations in one or several genes with clinical potential. Ultimately, they provided lists of potential targets for therapeutic development. Several LSDBs also include expert-curated, gene-specific information. The TP53 mutation database includes functional and structural information that led to the accurate classification of the loss of function of the various TP53 variants, information that is more specific than that provided by global algorithms such as SIFT or MUTAssessor (Reva et al., 2011; Sim et al., 2012). Curated mutations in LSDBs are also used as training or test sets in the development of novel biocomputing tools used for the stratification of driver and passenger mutations. The TP53 database was recently curated using specific statistical tools, resulting in the removal of more than 100 articles reporting artifactual data [Edlund et al., 2012] and Leroy et al. in the same issue. This curated database was used for the development of CHASM, a program for estimating the impact of missense mutations [Cline and Karchin, 2011].

Although cancer gene LSDBs contain mostly somatic mutations, germline mutations in several tumor suppressor genes associated with cancer predisposition are also available. Indeed, mostly germline mutations have been described for several genes, for example, *BRCA1* or *BRCA2* (MIM #600185) associated with breast and ovarian cancer, or *MSH2* (MIM #609309) and *MLH1* (MIM #120436) associated with hereditary nonpolyposis colorectal cancer, whereas both germline and somatic mutations are available for other genes, for example, *APC* (MIM #611731) associated with familial adenomatous polyposis or TP53 associated with Li–Fraumeni syndrome (see the review of Kamihara et al. in this issue for a full discussion on TP53 germline mutations).

Although the pathogenicity of somatic mutations in cancer genes can be easily appraised, assessing the pathogenicity of germline mutations is far more difficult. Providing patient counseling is critical in the presence of a suspected hereditary syndrome. LSDBs have been invaluable as tools for classifying these variants and as a reference for clinical geneticist.

We can nonetheless predict that LSDBs, in their current form, will disappear very quickly. The reasons for this include the lack of mutation descriptions in publications, information redundancy with large databases and an absence of funding. Indeed, since the launch of the various large tumor-sequencing projects, LSDBs on novel cancer genes are no longer being developed. Furthermore, cancer cannot be reduced to the simple presence of an SNV; other genetic and epigenetic events participate actively in the tumorigenesis process. The versatility of novel NGS platforms allows for multiple types of analyses, including sequencing, copy number evaluation, translocation detection, methylation profiling, and expression profiling. By combining these analyses, researchers will be able to draw an integrated picture of each tumor and pinpoint relationships (or absence of relationships) between various types of alterations and specific pathways. The TCGA and ICGC databases are informationally complete and can be browsed via specific Web portals.

Nevertheless, a curated repository for each cancer gene must be available so that the scientific community has access to the same, accurate, bulk information previously supplied by LSDBs. As the number of true cancer genes will most likely be limited, it is not unimaginable that administrators of large cancer databases assign a group of curators/experts to each gene of importance. Specific information relevant to individual genes could be added to the database and made available to the community via specialized but easily accessible gene-specific interfaces. For example, The TP53 database includes more than 100,000 entries concerning functional data (e.g., transactivation, DNA binding, gain of function) for 2,000 TP53 variants (Leroy et al., 2013). In the *APC* tumor suppressor gene, the distribution of its biallelic nsSNV events is not random and knowledge of this distribution will be important for explaining the mutation spectra observed in colorectal cancer (Fearnhead et al., 2001). Furthermore, the localization of mutations in the *APC* gene is highly correlated with disease severity and association with extracolonic features.

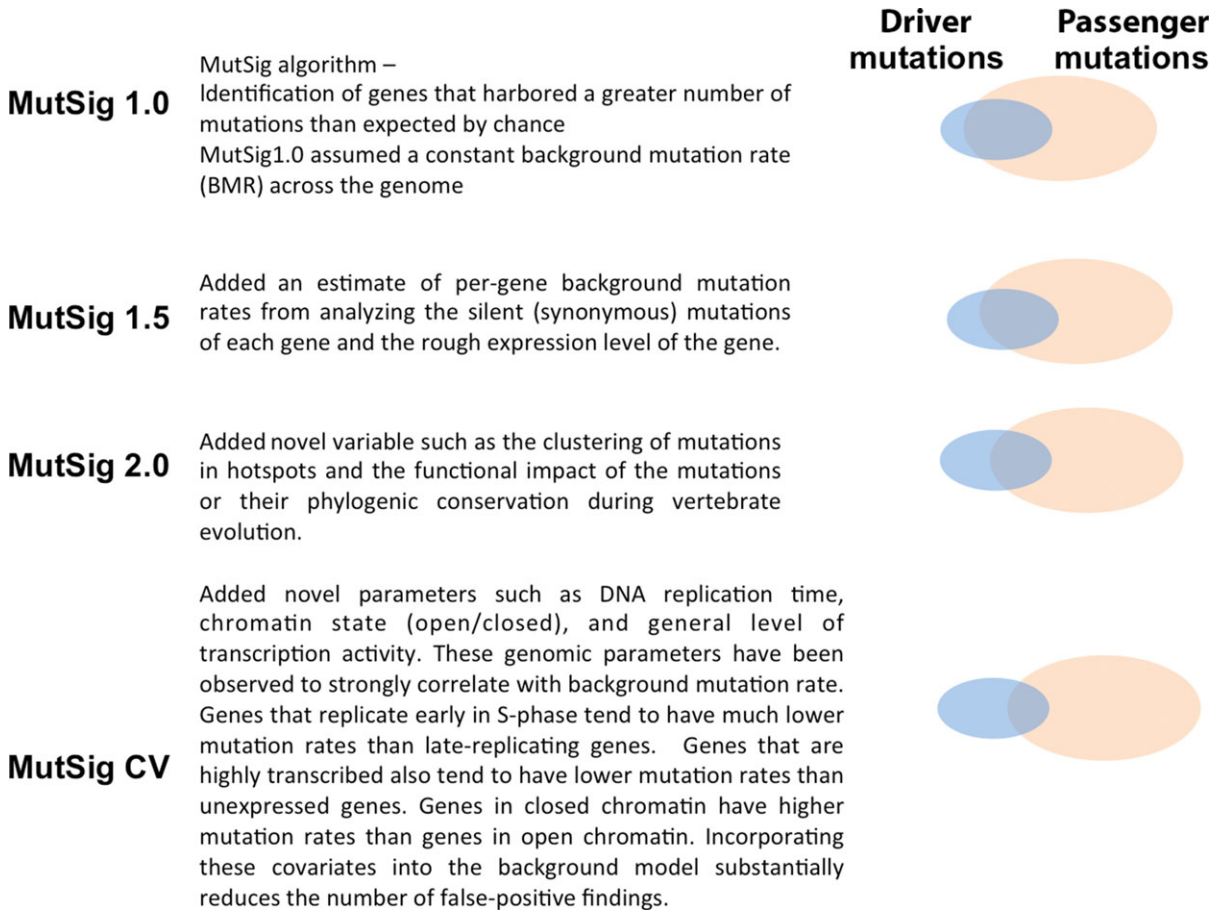

Figure 6. Development of the MutSig algorithm: Since the first description of MutSig in 2007, the algorithm has quickly evolved to better distinguish driver and passenger mutations. In the first version, spurious genes such as TTN, MUC4 or MUC16 were defined as potential driver mutations in various types of cancer. However, MutSig evolved across several versions to ultimately provide better discrimination and eliminate many false positives in the current version, as described by Lawrence et al. (2013) MutSigCV includes multiple biological parameters that strongly influence the rate of mutation, such as replication timing and transcriptional activity. As these two parameters are tissue specific, future versions of MutSig will have to include the type of cancer. The increasing availability of data via larger sequencing projects (e.g., the 10K project discussed above) will provide more power to improve the accuracy of this algorithm. Figure provided through the generosity of Broad Institute, Inc. and adapted from http://www.broadinstitute.org/cancer/cga/mutsig. © 2013 Broad Institute, Inc. all rights reserved.

Ultimately, what should be the most desirable output? For a clinical oncologist dealing with patients or families on a daily basis, reports of genetic modifications must be linked to the full description of their biological consequences so that the information can be used with strong confidence. The number of allelic variants of unknown significance, a nightmare for clinical geneticists, must be reduced to lessen the burden for both clinicians and families. The integrated association of the multiple mutations found in a tumor and the genetic background of the patient will help to define the most effective therapy.

For basic research, as discussed above, disease-associated mutations are of inestimable value. Gene deconstruction via disease-associated mutations can be likened to evolution but it shapes the functionality of the gene product to another purpose, such as hyperactivity, partial or total loss of activity or the gain of a new function. The propensity of a specific nucleotide to be selected during the neoplastic transformation is related to its mutability depending not only on the DNA sequence context, but also and more importantly on the consequences of the alteration at the RNA and/or protein level. Therefore, each residue of a protein can be associated with a decomposition factor in contrast to evolutionary conservation and cancer hot spot mutation residues associated with a high decomposition factor are frequently localized at phylogenetically conserved positions. The value of cold spots for mutation, that is, residues highly conserved phylogenetically and never selected in cancer, have been largely underestimated. In the TP53 gene, only a few conserved residues of the protein are never found to be mutated in human cancer. It appears that these residues control the interaction of the negative regulator mdm2 with TP53, and any alteration would be lethal for the cell (see Leroy et al. in this issue for more information). This information gained via the analysis of the mutation database in association with a phylogenetic analysis, the construction of artificial mutants and structural analysis of the TP53 protein, demonstrates to what extent information dispersed in various databases or publications can raise pertinent questions.

Connecting all the information stored across various databases in such a way as to make it accessible via a single output interface will be one of the most important challenges to meet in the coming decades. Indeed such an orchestrated approach will help calm the sea of big data and thus allow the scientific community to efficiently use this information in its quest to understand and treat cancer.

Acknowledgments

Our work is supported by Cancerföreningen i Stockholm, Cancerfonden, and the Swedish Research Council (VR).

Disclosure statement: The authors declare no conflict of interest.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, et al. 2013. Signatures of mutational processes in human cancer. Nature 500:415–421.

Auerbach AD, Burn J, Cassiman JJ, Claustres M, Cotton RG, Cutting G, den Dunnen JT, El-Ruby M, Vargas AF, Greenblatt MS, Macrae F, Matsubara Y, et al. 2011. Mutation (variation) databases and registries: a rationale for coordination of efforts. Nat Rev Genet 12:881; discussion 881.

Baker SJ, Fearon ER, Nigro JM, Hamilton SR, Preisinger AC, Jessup JM, vanTuinen P, Ledbetter DH, Barker DF, Nakamura Y, et al. 1989. Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. Science 244:217–221.

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, Cortes ML, Fernandez-Lopez JC, et al. 2012. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature 486:405–409.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Biesecker LG, Spinner NB. 2013. A genomic view of mosaicism and human disease. Nat Rev Genet 14:307–320.

Bond GL, Hu W, Bond EE, Robins H, Lutzker SG, Arva NC, Bargonetti J, Bartel F, Taubert H, Wuerl P, Onel K, Yip L, et al. 2004. A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. Cell 119:591–602.

Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, Stratton MR. 2008. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci USA 105:13081–13086.

Chanock SJ, Thomas G. 2007. The devil is in the DNA. Nat Genet 39:283–284.

Claustres M, Horaitis O, Vanevski M, Cotton RG. 2002. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 12: 680–688.

Cline MS, Karchin R. 2011. Using bioinformatics to predict the functional impact of SNVs. Bioinformatics 27:441–448.

Collins FS. 1995. Positional cloning moves from perditional to traditional. Nat Genet 9:347–350.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12:628–640.

Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, et al. 2002. Mutations of the BRAF gene in human cancer. Nature 417:949–954.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025.

Edlund K, Larsson O, Ameur A, Bunikis I, Gyllensten U, Leroy B, Sundstrom M, Micke P, Botling J, Soussi T. 2012. Data-driven unbiased curation of the TP53 tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors. Proc Natl Acad Sci USA 109:9551–9556.

Fearnhead NS, Britton MP, Bodmer WF. 2001. The ABC of APC. Hum Mol Genet 10:721–733.

Garraway LA, Lander ES. 2013. Lessons from the cancer genome. Cell 153:17–37.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, et al. 2012. Architecture of the human regulatory network derived from ENCODE data. Nature 489:91–100.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862–864.

Gupta R, Webb-Myers R, Flanagan S, Buckland ME. 2011. Isocitrate dehydrogenase mutations in diffuse gliomas: clinical and aetiological implications. J Clin Pathol 64:835–844.

Guruharsha KG, Kankel MW, Artavanis-Tsakonas S. 2012. The Notch signalling system: recent insights into the complexity of a conserved pathway. Nat Rev Genet 13:654–666.

Hawkins RD, Hon GC, Ren B. 2010. Next-generation genomics: an integrative approach. Nat Rev Genet 11:476–486.

Hood L, Rowen L. 2013. The human genome project: big science transforms biology and medicine. Genome Med 5:79.

Horaitis O, Cotton RG. 2004. The challenge of documenting mutation across the genome: the human genome variation society approach. Hum Mutat 23:447–452.

Horn S, Figl A, Rachakonda PS, Fischer C, Sucker A, Gast A, Kadel S, Moll I, Nagore E, Hemminki K, Schadendorf D, Kumar R. 2013. TERT promoter mutations in familial and sporadic melanoma. Science 339:959–961.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, Hill DP, Kania R, Schaeffer M, St Pierre S, Twigger S, White O, et al. 2008. Big data: The future of biocuration. Nature 455:47–50.

Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. 2013. Highly recurrent TERT promoter mutations in human melanoma. Science 339:957–959.

Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, et al. 2010. International network of cancer genome projects. Nature 464:993–998.

Ingram VM. 1956. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. Nature 178:792–794.

Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, Hutchinson A, Deng X, Liu C, Horner MJ, Cullen M, Epstein CG, et al. 2012. Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet 44:651–658.

Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael JF, Wyczalkowski MA, Leiserson MD, Miller CA, et al. 2013. Mutational landscape and significance across 12 major cancer types. Nature 502:333–339.

Kern SE, Winter JM. 2006. Elegance, silence and nonsense in the mutations literature for solid tumors. Cancer Biol Ther 5:349–359.

Kern SE, Kinzler KW, Bruskin A, Jarosz D, Friedman P, Prives C, Vogelstein B. 1991. Identification of p53 as a sequence-specific DNA-binding protein. Science 252:1708–1711.

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, Sboner A, Lochovsky L, Chen J, Harmanci A, Das J, Abyzov A, et al. 2013. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342: 1235587.

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The next-generation sequencing revolution and its impact on genomics. Cell 155:27–38.

Kohonen-Corish MR, Al-Aama JY, Auerbach AD, Axton M, Barash CI, Bernstein I, Beroud C, Burn J, Cunningham F, Cutting GR, den Dunnen JT, Greenblatt MS, et al. 2010. How to catch all those mutations–the report of the third Human Variome Project Meeting, UNESCO Paris, May 2010. Hum Mutat 31:1374–1381.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214–218.

Lee W, Yue P, Zhang Z. 2009. Analytical methods for inferring functional effects of single base pair substitutions in human cancers. Hum Genet 126:481–498.

Leroy B, Fournier JL, Ishioka C, Monti P, Inga A, Fronza G, Soussi T. 2013. The TP53 website: an integrative resource centre for the TP53 mutation database and TP53 mutant analysis. Nucleic Acids Res 41:D962–D969.

Li Y, Gordon MW, Xu-Monette ZY, Visco C, Tzankov A, Zou D, Qiu L, Montes-Moreno S, Dybkaer K, Orazi A, Zu Y, Bhagat G, et al. 2013. Single nucleotide variation in the TP53 3' untranslated region in diffuse large B-cell lymphoma treated with rituximab-CHOP: a report from the International DLBCL Rituximab-CHOP Consortium Program. Blood 121:4529–4540.

McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. 2013. Impact of deleterious passenger mutations on cancer progression. Proc Natl Acad Sci USA 110:2910–2915.

Metzker ML. 2010. Sequencing technologies—the next generation. Nat Rev Genet 11:31–46.

Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11:685–696.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80.

Ozsolak F, Milos PM. 2011. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12:87–98.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680.

Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL, Olivi A, McLendon R, et al. 2008. An integrated genomic analysis of human glioblastoma multiforme. Science 321:1807–1812.

Pennisi E. 2012. Genomics. ENCODE project writes eulogy for junk DNA. Science 337:1159, 1161.

Reva B, Antipin Y, Sander C. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39:e118.

Sanderson K. 2011. Bioinformatics: curation generation. Nature 470:295–296.

Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12:683–691.

Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. 2010. Computational solutions to large-scale data management and analysis. Nat Rev Genet 11:647–657.

Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40:W452–W457.

Soussi T, Caron de Fromentel C, Mechali M, May P, Kress M. 1987. Cloning and characterization of a cDNA from Xenopus laevis coding for a protein homologous to human and murine p53. Oncogene 1:71–778.

Soussi T, Beroud C. 2001. Assessing TP53 status in human tumours to evaluate clinical outcome. Nat Rev Cancer 1:233–240.

Soussi T, Ishioka C, Claustres M, Beroud C. 2006. Locus-specific mutation databases: pitfalls and good practice based on the p53 experience. Nat Rev Cancer 6:83–90.

Stratton MR. 2011. Exploring the genomes of cancer cells: progress and promise. Science 331:1553–1558.

Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature 458:719–724.

Susumo O. 1972. So much "junk" DNA in our genome. In: Smith HH, editor. *Evolution of genetic systems*. New York: Gordon and Breach. p 366–370.

Takahashi T, Nau MM, Chiba I, Birrer MJ, Rosenberg RK, Vinocour M, Levitt M, Pass H, Gazdar AF, Minna JD. 1989. p53: a frequent target for genetic abnormalities in lung cancer. Science 246:491–494.

Tomso DJ, Inga A, Menendez D, Pittman GS, Campbell MR, Storici F, Bell DA, Resnick MA. 2005. Functionally distinct polymorphic sequences in the human genome that are targets for p53 transactivation. Proc Natl Acad Sci USA 102: 6431–6436.

Varley JM, Chapman P, McGown G, Thorncroft M, White GR, Greaves MJ, Scott D, Spreadborough A, Tricker KJ, Birch JM, Evans DG, Reddel R, et al. 1998. Genetic and functional studies of a germline TP53 splicing mutation in a Li–Fraumeni-like family. Oncogene 16:3291–3298.

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LAJ, Kinzler KW. 2013. Cancer genome landscapes. Science 339:1546–1558.

Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S, Stemke-Hale K, Davies MA, Gershenwald JE, Robinson W, Robinson S, et al. 2011. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet 43:442–446.

Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, et al. 2007. The genomic landscapes of human breast and colorectal cancers. Science 318:1108–1113.

Yates LA, Norbury CJ, Gilbert RJ. 2013. The long and short of microRNA. Cell 153:516–519.

Yates LR, Campbell PJ. 2012. Evolution of the cancer genome. Nat Rev Genet 13:795–806.

Zeron-Medina J, Wang X, Repapi E, Campbell MR, Su D, Castro-Giner F, Davies B, Peterse EF, Sacilotto N, Walker GJ, Terzian T, Tomlinson IP, et al. 2013. A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection. Cell 155:410–422.