Discrepancies in database-dependent research and proactive management of project procedures and structure to adapt

Manousos E. Kambouris [1,2]

(1) The Cancer Institute of New Jersey and the Department of Molecular Genetics, Microbiology and Immunology, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey 08903, USA
(2) Department of Medical Laboratories, Faculty of Health and Caring Professions, Technological Educational Institute of Athens, Ag Spyridonos st., 12210 Athens, GREECE.

Abstract

The need of public access databases in modern biomedical research and to other disciplines is well-attested, provided that they are easily accessible, user-friendly, compiled with quality critaeria ensuring and bolstering their use and the data provided are dependable, straightforward in meaning and presented in usable form (effect-oriented data bases). The organization, maintenance and presentation of such databases are expensive and laborious tasks, thus encompassing significant funding. Their promotion and use, though, should follow some standards and rules and be coupled with compatible research project structure, reviewing and granting, a combination which will allow fruitful use of databases' resources and help avoid previous misconceptions in the management and allocation of the related research resources. The objective is to avoid the pitfalls of the past and to further the trustworthiness of the data provided early on the creation of a database which by definition is not only limited but also of poorer quality and limited uniformity, and structure research projects accordingly so as to allow limited but reliable use of such resources as early as possible with little risk for the user.

Keywords: The Human Genome Project, SNP, database, project management, project structure

Introduction

The compilation, expansion and widespread availability of effect-oriented (as opposed to knowledge-oriented[1]) databases launched the biomedical sciences into a new age just before the turn of the millennium (8). A new generation of disciplines, the "-omics", based on extensive data-mining and holistic approaches to either quickly tackle a wide research or to dissect the peculiarities of complex phenomena (4, 5, 6, 7, 9) was brought forward as a direct, and probably deliberate result of this strategic choice in science management and financing. Such databases needed massive, multi-central efforts and substantial funding to materialize, whereas the managing and conditioning of many different contributors from throughout the globe was by itself a tremendous task (2, 5). The funding required was quite substantial and was largely seen as an investment which would pay back by the thrust it would provide to subsequent research, both basic and applied (7). Naturally, research projects tailored to making use (any use) of these resources were receiving priority for grants (1, 10)

[1] Simple compilation of relative knowledge packages without criteria and standards promoting or even permitting their use from independent users

especially whenever the granting authority had invested or contributed to the creation of such databases. The reason was two-fold:

First, using the database resources was excellent Public Relations (PR) for contributors and investors /fund providers. Not only the usefulness of the database project were made obvious, which automatically enhanced the stakes involved, but further funding for expansion, organization and evolution of the databases could be attracted in a self-supported way. This in part referred the upfront costs to a wider base and allowed the database compilation and construction to evolve, multiply and propagate to other fields and applications, cascading the investment (7).

The second reason was that such use of compiled knowledge provided a degree of both real and PR-related security in that solid background was at hand for even the most ambitious projects. The latter were deemed high-risk, high-payoff efforts, but the regulation of the risk due to the massive knowledge assembled helped a great number of participants to apply. So, as the databases were still evolving through the funding attracted by their perceived usefulness (7), in a more or less spiral development, the database-inspired projects allowed optimism for multiple, quick and extensive breakthroughs in knowledge, technology and applications (4, 5, 10).

Unfortunately since the databases were still **morphing** and their usefulness was essential for their consolidation, rather than the opposite (database usefulness stemming from its quality and stability), management problems and interaction jams appeared in a massive scale. Such was the case with the Human Genome Project and spin-offs of its basic data compilation, as were the Single Nucleotide Polymorphism (SNP) maps, which rippled research practices after the turn of the Millennium.

A case study

The dynamic nature of the public SNP database of "The SNP Consortium" (TSC)[2], evolving in parallel with the Human Genome Project (HGP) represented in mid-2000 a challenge to cope with for many a project based on the establishment of physical SNP maps to evaluate correlation of genetic loci. This drawback could by definition stall -or even ground altogether- a project, regardless the latter either being focused on a specific, highly interesting chromosome or chromosomal region (such as Chromosomes 1 and 14 and the p- region of Chromosome 17 or parts of Chromosme X respectively) (1, 3, 4) or extending to a genomewide analysis project (10) in search for interrelated genes and other genetic loci. The most optimistic estimations were that it would take at least a year for a more stable condition to be achieved. Such a timeframe was almost unacceptable to many labs and teams which had been charged with non-recurring costs of recruitment and hardware investments; the specific software products, necessary in most of the cases for handling such studies were also falling back in schedule, bringing in more discomfort and economic burden to the -anyway limited- budgets. Moreover, the strategy followed by The SNP Consortium (TSC), conceived to enable progress in parallel with the HGP (2), saved time compared to a tandem pursuit of the two projects, but made rather improbable the lapse of less than 3 years for a complete, stable, reliable and in-depth informative version of the database with integral access to population data[3]. Such prospects furthered the horizon of SNP-related projects to a 5-year frame, if the need for

[2] Formed mostly by private sector entities (2, 7)

[3] Actually, the HGP was declared complete on April 2003 and this was an absolute prerequisite for any interrelated database to consolidate.

development of experimental procedures were taken into account. This invariably meant that the acquired or envisaged equipment would largely be obsolete, both in hardware and software, before having the chance to be used for the projects that granted its acquisition. This is especially true for leading edge technologies like laser reading and scanning systems, arrayers and automatic sequencers with capillary or other advanced formats of the day (2, 5, 10).

The best – if not the only- way of circumventing such depressing prospects was by organizing the experimental tasks and procedures as to evolve in parallel with the ongoing project of TSC. Such a solution, however, has some limitations: It practically means that the specific means earmarked for use with the database for data mining are to remain inert or, in the best case scenario, sub- used for other tasks. Also, many of the experimenting developmental procedures (especially the SNP-independent, such as electrophoretic and amplification procedures) were definitely due to overcome any obstacles and be completed far earlier than the abovementioned timeframes, whereas SNP-related methodologies were to face a deadlock when reaching the stage of mandatory fusion of collected SNP data (3). A drastic delay in the latter case was most likely to occur and to lead to premature obsolescence the whole methodology, either imposing a great financial burden for replacement/upgrade, or becoming a limiting factor for future progress, with far-reaching consequences of every kind, since the obsolete methodology is to be used, although not optimized, at least to the point of the initial investment being paid off.

A way to overcome this prospect is by using random, low quality, even virtual SNP data only as simulation input in order to accomplish the development stage, and later to incorporate the true SNP data when they become available. But in some cases the whole procedure is designed as a function of the exact nature of the true data and a separation of the procedures in terms of time could well nullify the whole project, or, in other, simpler cases, cause a need for project restructure, a procedure of proven burdensomeness in terms of funds, labor, and time.

In such a context, a set of rules of use for data mining into the database were proposed. These cannot ensure the accuracy or validity of the selected SNP data, but could help avoid at least the need for repeated wholescale editing, and also present the easiest way to edit and update the collected data in due time. In this way, rigidly organized and structured projects could well progress with less uncertainty and limited need for extended revising caused by the almost continuous (weekly) update of the TSC database.

Two distinct trends have been observed that resulted in intrachromosomal relocation of SNPs and contigs[4]. One trend is the shrinkage of the physical chromosome size, which leads to contigs' slipping away of their original locations, usually to move closer to one another. This trend, however, is not universal and many subsequent builds re-elongated, up to a point, the chromosomes and subsequently moved the SNPs further away from each other. The second trend is that the stablest part of the chromosomes seemed to be the q-telomer, which was well documented as far as both contigs and SNPs were considered since early Dec 2000.

[4] Contigs: Contiguous blocks of DNA compiled from the alignment of partially overlapping sequenced DNA fragments, produced by cloning of restriction fragments of the original sample chromosome. Contigs are superimposed onto other chromosomic (i.e. cytarogenetic) maps through the use of tags and marker sequences.

These two observations indicated that in order to limit the extend of relocation to be caused by any ongoing or subsequent shift of chromosome size and resulting positional changes, the wisest course to proceed in SNP selection would be: a) to take as a reference point the q-telomer and b) establish a simple, purely arbitrary size scale, which begins from the q-telomer and progresses towards the p-telomer, being inverse in direction compared to usual chromosomal studies. Such a course of action might have limited the relocation effort needed in subsequent amendments, despite the widespread, over-optimistic belief that the drastic slippage of contigs along the chromosomes were over since winter 2000-1).

This, "inverse" arbitrary system had another direct advantage: It permits starting from a well-established area, with fewer problems, and then extend towards more problematic areas. This is especially true with acrosomatic chromosomes (i.e. 21, 22 etc) in which case there was no easy way to establish where exactly the unsequenced, highly repetitive and problematic parts end and thus where exactly the mapped sequences begin. By an inverted distance determination system, a more orthodox approach would ultimately be possible, where the beginning is securely anchored at the q- telomer and each pace of progress towards the unidentified areas of the acrosome is simply to elongate constantly the known sequences, instead of having to wheel back the whole part of known sequences every now and then. This reverse procedure was also to help accurately determine the then-exact location of the border between the already chartered and the problematic parts of such chromosomes and readily identify any progress in the future by simply watching the updated score. The exact knowledge of the physical location of the border between chartered and unchartered regions (relatively to a fixed and rather stable reference point as the q-telomer seems to be) would also assist in our defining the true physical size of the problematic regions, thus facilitating any effort to chart these regions in the near future as well.

Unfortunately there is simply no way known to us to safeguard to any extend from SNPs "hopping". This term means their relocation at entirely different chromosomal regions, in some cases Mbps away from their originally acknowledged location, or even onto different contigs and chromosomes. The removal of the actual mapped database of some thousands of SNPs in certain updates just points out how probable such an occasion is always going to be, especially since TSC might use more and more stringent alignment criteria, as the HGP progresses. More stringent criteria were deemed an absolute necessity, as they reduce the number of possible alignment matches throughout the genome and thus render valid a great number of observed SNPs, previously rejected on the grounds of multiple alignments (>3) which could signify location on repetitive regions. It must be remembered that SNPs were not identified upon sequencing of contigs, but identified individually and independently and then blasted to sequences determined by the HGP. The only way to reduce the probability of seeing a selected SNP, in the future, hopping some tens of Mbps away, or onto another chromosome, would be to select, for whatever the purpose is, SNPs submitted with large flanking sequences, as such a quality makes more reliable the blasting and alignment result.

One of the most prominent dangers as already said was the irregularly emergence of "new" (newly discovered or newly aligned) SNPs in the whereabouts of existing ones. Some applications might well not be affected, but others, as are the design of PCR (Polymerase Chain Reaction) and SBE (Single Base Extension) primers, definitely are (2, 10). Statistically, it could be maintained that selection of SNPs in chromosomal regions where their number and density is low could offer a higher degree of safety against such prospects, compared to regions flooded with SNPs. Naturally, this course of action cannot be maintained in medium- and high-density mapping efforts (1, 4), especially when the objective is a map with standardized physical SNP distances.

Moreover, things could well prove more complicated than simple statistics might imply, based on the observed density divergence into the same chromosome (the SNP density greatly varied –up to 200-fold - among regions of the same chromosome and this has been reported for most, if not all, the chromosomes). So, there was a distinct possibility that some of the then-poor SNP-wise regions presented a rather deceptive picture, for two different reasons: The first reason is that such regions had been problematic in alignment and sequencing, and thus SNPs might exist in them in great numbers but either had not yet been discovered, or, more probably, had not yet been allocated to them due to blasting and alignment difficulties. Ironically, a good reason for blasting and alignment problems could well have been the existence of many SNPs of medium to high heterozygocity ratios, which would result in disqualification during the alignment procedure with more stringent algorithms and criteria. This means there is a possibility that the regions which at a time appear as poor in SNP incidence might be in fact too rich in SNPs for them to be identified and displayed with the means and tools of the day.

The second reason is stemming from the heterozygosity ratios. Up to-date the most well -documented SNP submissions were generally limited to a sample size of 10 and a great number of submissions was restricted to a sample number of merely 4 and in many cases just 2. This practically means that, in the best of cases, SNPs of down to 10% heterozygocity ratio were detectable and thus scored, but in many cases scoring was limited to 25% or 50% ratios. This implies that areas (of variable sizes) deemed SNP-free, might in truth house SNPs of low heterozygocity ratios and even be of remarkable SNP density.

As TSC proceeded with more samples for confirming existing SNP and for providing adequate heterozygocity data, this problem was prone to occur acutely, and to get even be more perplexed by two other factors: First, population differentiations in SNP heterozygocity ratios. This factor could well be as severe as becoming qualitative and not quantitative, if in some populations a SNP site had been represented only in one allelic form and thus be deemed homozygous. Second, there had been immense discussion as to what exactly should be deemed as "true" SNP. Technical discrepancies notwithstanding, still there was the problem of the genetic stability of a "look-like- a- SNP" to be judged as such. The latter issue is even more embarrassing, as the stability in such contexts is really a matter of cell generations and our techniques are focused onto human individuals' generations. It is very obvious that in some cases human cell generations are almost a match for human individual ones, but in other cases, as in cells and tissues reproduced throughout the human lifetime, things are very different. And, of course, if the unforetold emergence of previously non-existent unknown SNPs could mar a SNP-based project, the deletion of selected SNPs was even worse. Although this possibility was deemed negligible at the time, especially compared with the possibility of relocation or "hopping", the notification posted on 30/9/08[5] in the database proper warned the "SNP users" for a considerable decrease in the number of SNPs at the most recent builds of the database.

[5] **Pasted on 7 Aug 2009**
"Attention dbSNP user:
We discovered two problems with SNP annotation on RefSeq mRNA.
Problem 1: A drop in the total number of SNP annotations from dbSNP build 129 onto human mRNA sequences for RefSeq

Thus, nobody had really been in a position to make any substantiated proposals as to how SNPs should at the time be selected in order to minimize the prospect of being found among a multitude of other SNPs popping up later. A thumb rule could be, despite what was mentioned earlier, to prefer rather scarce regions when there was a choice. Even more important might prove a concerted effort in the methodologies' developmental criteria that could allow for as short as possible definite flanking sequences. In this scheme, one should ideally allow as lenient as possible matching criteria regarding as short as possible flanking sequences and at the same time inspect only SNPs submitted with as long as possible flanking sequences, which by definition yield a higher positional reliability.

Although the above really offered little scope for positive insurance against future rearrangements, there has been a definite factor able to ensure a fair degree of negative insurance and serve as a thumb rule for utter avoidance: when selecting SNPs from the SNP database, one is in many cases encountered with multiple submissions under the same Reference Sequence Number (rs#). All separate submission pages, even if coming from the same lab/submitter, must be opened and looked at, to avoid the possibility of two different submissions under the same rs number providing two actually different SNPs. This has been the case many a time and the dbSNP (data base SNP) staff issued a relevant warning in the SNP page of NCBI database since 30/9/08 (see footnote 2). Such inconsistency can be attributed to the automated methods used by some submitters and the strategy for aligning the flanking sequences. The independent alignment of the flanking sequences of each SNP might well result in two, different but close-by SNPs to be simultaneously attributed to the same, confined area of the genome and thus be scored by the automated systems as the same, under one SNP rs number. However, and not regarding the reasoning of the event, the significance of the issue is clear-cut: A second, low-heterozygocity SNP position is in the immediate vicinity of another, "dominant" one and this sould be taken into consideration when the experimental procedures to be developed might be hampered by such incidents.

On the other hand, the multiple, independent submissions, in theory, when not contradicting, had a beneficial spin-off: As most of the laboratories generally submited a SNP after examining a steady sample size, the individual sample sizes could be deduced and a more accurate idea of the heterozygocity ratio could be predicted early on, even if not mnentioned. For example, if the total sample size were 12 chromosomes, and there were two submissions, one of a lab that used to submit SNPs detected in 10- chromosome samples and another from a lab that submitted SNPs detected in either 2- or 4-chromosome samples, it is easy to deduce that in this specific case the latter lab submitted a SNP from a 2-chromosome sample and thus the heterozygocity ratio is probably near 50%. The submission of the 10-chromosome sample by itself would offer enough reliability (compared to the total of 12 chromosomes sampled), but the heterozygocity implied could be everything between 50 and 10%. Unfortunately this approach was valid only when there were multiple

releases 28, 29, and 30.
More...
Problem 2: dbSNP has also identified a separate problem of redundant SNP annotations on RefSeq mRNAs where the same rs number is annotated more than once on the same mRNA.
More...
If you have specific questions regarding these problems please contact us at snp-admin@ncbi.nlm.nih.gov.
We apologize for the inconvenience that these errors may havecaused.
Best regards,
dbSNP Staff.
09/30/2008"

submissions from different labs, all really referring to the same SNP, with the same flanking sequences. In case one of the labs submitted the same SNP more than once, its multiple submissions might be caused by repetitive electronic function and thus should be considered a single entry. This only if the submitted sequences were consistent with each other and/or to identical submissions from other labs.

The main source of problems lied to the absence of any universal set of submission/acceptance criteria for the SNPs. Not taking into consideration the potentially extremely important difference of the human source/ donors the various labs use for the tracking of SNPs, the submission procedure by itself did much to make things less than perfect. Some labs reverted to submission with standard flanking sequence sizes. Sizes of 25, 35, 50 60 and 200 bases were noticed. Unfortunately, a flanking sequence size of less than 100 bases is utterly insufficient for primer design manipulations that would permit a really meaningful use of the submitted SNP, especially when the case is multiplex formats[6]. Labs that joined The SNP Consortium later on, seemed to have realized the need for uniformity, but not the need for “adequately long” flanking sequences.

Other labs did not keep a standard in the submission length. So, there were cases of “moderate” flanks, of a few hundred bases, cases of massive flanks of up to 10 thousand bases (which, by the way, frequently contain other SNP loci without, of course, marking them) and cases of greatly asymmetric flanks, where one flank is sometimes less than 25 bases and the other many tens to some hundreds of bases.

The standard of SNP quality is also non-existent. Submissions initially varied from 2 to 10 identical (from the same individual) or homologuous (from different individuals) chromosomes being tested. Later this number increased to 50 chromosomes, in some cases from a single lab, in other cases as a parallel effort of more than one labs. Still, usually the allelic frequencies were not mentioned- nor did the identical or homologuous status of the sample chromosomes. This is understandable, as for such data it is important to have a statistically significant sample size, which the 50 chromosomes are not, especially when picked from a variety of labs and a pattern of donors/subjects. On the other hand, not mentioning any information of allelic frequency made impossible any arbitrary selectivity in picking the SNPs for other studies. Thus, a SNP submitted after scanning 2 chromosomes bears a potential minor allele frequency of 50%, whereas a SNP submitted after scanning 50 chromosomes might bear a minor allele frequency of 2%. It is known, though, that the threshold between SNP and mutation was originally set to heterozygocity ratio of >1%. Although a SNP documented in 50 chromosomes has an inherently higher degree of reliability, if nothing else, from being a misalignment, artifact or pseudoSNP, the SNP documented in 2 chromosomes offers the intriguing and appealing prospect of higher minor allele frequency. This paradox was of extreme importance for mid-and short-term projects based on and supported for the collection of large numbers of SNP markers.

Lessons learned?

The chosen- and highly celebrated- “commercial approach” to the TSG created dire problems. The managers, boasting to interviews that they check the progress of the

[6] For such applications Tm considerations and primer reaction parameters must be satisfied simultaneously and thus longer flanking sequences, allowing more alternative primer sequence designs are favorable (3).

submitters daily, meaning exercising pressure to increase productivity, paid little attention to the quality and usefulness of the volume of the project. The strategic decision to jumpstart the exploitation of the database in order to maintain and increase funding, encouraging its use in projects and thus deferring part of the costs to the end users (scientific projects making use of the data and their sponsors) caused unknown as to yet damage to many labs that were encouraged to show trust to the product of the TSG and to rely on it for their projects, and subsequently failed to deliver the projected results to their sponsors, losing credibility and future funding. The politically correct language, describing the database as "dynamic in nature" or "fluid", instead of "unstable" or "unreliable" solved none of the massively emerging problems. Once commercial approaches are followed in scientific projects, one must expect both their good (motivation, flexibility, accountability) as well as their bad (increased reliance on PR, substandard products and services, projections instead of facts) aspects to emerge. The maturing of HGP database and all of its spinoffs would have allowed far better focused and substantiated projects with higher success rates and lower costs and risks.

When the decision to start capitalizing on the massive and prolonged investment of the HGP was made, it was well-known the extend of the progress. The nature of the project meant that usable information would not occur piecemeal, but in swarms. To fully exploit the first such surge of usable data, projections were made with specific targets, so as to lure investigators to include the database resources into their projects. But although of a massive, almost industrial scale and of routine methodology, the HGP had bottlenecks and pitfalls from the very beginning, where progress was uncertain and breakthroughs needed and not guaranteed. And these problems affected the progress of the SNP database, which was interlinked to the HGP. Betting on the best case scenario (2), the management tried to fully exploit the potential. But once the bet was lost, this approach caused multiple, cascading project failures with huge costs. For such far-going projects it is imperative to invest on and exploit confirmed, checked and quality-controlled products and services, not on projections and prediction. Spiral development, with successive builds and upgrades, produces an illusion of continuous, almost predictable progress but it is beneficial only if early versions cover some initial, time-critical needs. When the user's needs are far-fetching and demand the later versions of development, the spiral procedure brings no advantage whatsoever; practically, it is counter-productive since it slows the end result and defers money for integrate interim products. A part of early results might well be useful in some applications, but to truly exploit it with the least possible risk highly expedient management is needed, instead of aggressive one. If the actual results, being QCed and confirmed were publicized and made available, then far less projects would have been proposed, with less PR advances. Instead, the projects that did so would have better chances of success and would entail a far better implementation ratio and cost-effectiveness. This compilation and confirmation of the results of a continuously progressing effort is burdensome for the management, does injustice to its full potential since once completed it is already obsolete due to the fast pace of advancement, but secures the user from far-reaching and unpredictable pitfalls and creates reliability and trust. Encouragement to submit modular, flexibly managed projects based on generic, adaptable procedures is always advisable in such uncertain conditions, but rarely advised: in reality, it is generally avoided because such approaches increase the costs for a given target, whereas rigidly managed projects and highly integrated, optimized procedures allow for economies of scale and better cost-effectiveness.

References

1. Escary J-L, Bottius E et al (2000) “A first high-density map of 981 biallelic markers on human Chromosome 14”, *Genomics 70:* 153-64.
2. Hirschhorn JN, Sklar P et al (2000) “SBE-TAGS: An array-based method for efficient SNP genotyping” *PNAS 97: 22*, 12164-9.
3. Kambouris ME (2009) “Staged oligonucleotide design, compilation and quality control procedures for multiple SNP genotyping by Multiplex PCR and Single Base Extension Microarray format” *e-Journal of Science & Technology (e-jst), 4(2):* 21-40
4. Lai et al (1998) “A 4Mb high density SNP-based map around human APOE”, *Genomics 54:*31-8.
5. Lehnert V, Holzwart J et al (2001) “A semi-automated system for analysis and storage of SNPs” *Human Mutation 17:* 243-54.
6. Lockhart DJ &Winzeler EA (2000) “Genomics, gene expression and DNA arrays”. *Nature 405:* 827-36.
7. McCarthy JJ &Hilfiker R (2000) “The use of single nucleotide polymorphism maps in pharmacogenomics” *Nature Biotechnol 18* 505-8.
8. Stephens JC Schneider JA et al, (2001) “Haplotype variation and Linkage disequilibrium in 313 human genes”. www.sciencexpress.org/12 Jul 2001; 1-5
9. Velegraki A & Kambouris ME (2003) “Arrays and multiplex PCR: Revolutionary molecular biological methods with applications in biomedical practice” *Archives of Hellenic Medicine 20(4):*425–45
10. Wang HY, Luo M et al (2005) “A genotyping system capable of simultaneously analyzing >1000 single nucleotide polymorphisms in a haploid genome”. *Genome Research 15:* 276-83

Ασυμβατότητες στην έρευνα μέσω βάσεων δεδομένων και προσαρμογή προγραμμάτων με προληπτική διαχείριση διαδικασιών και διάρθρωσης

Μανούσος Εμμ Καμπούρης (1,2)

(1) Ινστιτούτο Καρκίνου Νέας Υερσέης και Τμήμα Μοριακής Γενετικής,
Μικροβιολογίας και Ανοσολογίας, Ιατρική Σχολή Robert Wood Johnson, Ιατρικό και
Οδοντιατρικό Πανεπιστήμιο Νέας Υερσέης,
New Brunswick, New Jersey 08903, USA
(2) Τμήμα Ιατρικών Εργαστηρίων, Σχολή Επαγγελμάτων Υγείας-Πρόνοιας,
ΑΤΕΙ Αθηνών, Αγ. Σπυρίδωνος, Αιγάλεω 12210

Περίληψη

Η χρησιμότητα των βάσεων δεδομένων στη σημερινή βιοϊατρική-και όχι μόνο-έρευνα είναι δεδομένη. Η οργάνωση, διατήρηση και διάθεση τέτοιων πόρων είναι πολυέξοδη και κοπιώδης και συνεπώς απαιτεί γενναιόδωρη χρηματοδότηση. Παρά ταύτα, η αξιοποίηση και χρήση τους πρέπει να διέπεται από κανόνες και πρότυπα και να συνδυάζεται με συμβατές δομές διαχείρισης, διάρθρωσης, αξιολόγησης και χρηματοδότησης ερευνητικών προγραμμάτων ώστε ο συνδυασμός να εξασφαλίζει επωφελή χρήση των βάσεων δεδομένων και να αποφεύγονται τα καταστροφικά σφάλματα του πρόσφατου παρελθόντος στη διαχείριση και των καταμερισμό των αντίσοτιχων κονδυλίων έρευνας.

Λέξεις-κλειδιά: Πρόγραμμα Ανθρώπινου Γονιδιώματος, Μονονουκλεοτιδικοί πολυμορφισμοί, Βάση δεδομένων, διαχείριση προγράμματος, διάρθρωση προγράμματος