Staged oligonucleotide design, compilation and quality control procedures for multiple SNP genotyping by Multiplex PCR and Single Base Extension Microarray format

Manousos E. Kambouris[1,2]

(1) The Cancer Institute of New Jersey and 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) Dept of Medical Laboratories, Fcty of Health and Caring Professions, Technological Educational Institute of Athens, Ag Spyridonos st., 12210 Athens, GREECE

Abstract

The value of the SNPs (Single Nucleotide Polymorphisms) as markers in various studies depends on their absolute numbers, but also on the more complicated issue of their availability. For any given study there must be enough SNPs suitably placed and satisfying specific generic qualitative criteria. The feasibility of a study depends on the ability to replace and keep the number of marker loci reasonably stable at each phase of the selected experimental procedures during the development of the overall methodology. We embarked on an experimental simulation, where we specified a series of criteria for the number and qualities of SNPs needed for a hypothetical study and proceeded to SNP selection and computerized primer design for each SNP locus. The decisive factor was the applicability of the primers in multiplex PCR format. The editing steps and software upgrades of the adopted selection procedure resulted in 98.6% primer efficiency. A total of 148 primer pairs were designed fulfilling multiplex PCR specifications, which were also effective in simplex quality control PCR assays against a start-up number of 150.

Keywords
oligonucleotide design, Multiplex PCR, Single Base Extension, DNA microarray, human genomic loci, QC assaying

Introduction

The launching, expansion and maturation of the NCBI SNP database has been hailed as a monumental project that will permit the use of SNPs as genetic markers for a variety of studies **(8)**. Among a multitude of studies where the SNPs might play a pivotal role, the ones that are expected to have the most impact are the genome-wide scans and other less restricted (intergenomic), or less expanded (single chromosome) scanning concepts of genomic nature **(3), (4), (10)**.

The data provided by The Snp Consortium (TSC) concerning the total SNP number throughout the human genome and the mean linear SNP density (number of total SNPs by total genome sequence length in kilobase pairs –SNPs/kbp) are rather optimistic figures regarding the practicality of using SNPs even for high-density mapping **(9)**. There are, though, some reasons for concern. The SNP dispersion varies highly in different chromosomes and chromosomal parts **(11)**. This means areas practically barren of SNPs and others so “populous” that they might be impractical for use due to their closeness (**Kambouris & Li, unpublished results**). Moreover, the use of SNPs as genetic markers must fulfill some parameters of paramount importance. A very important parameter is the actual location of a SNP. The use of markers presupposes "suitable" location for the best possible study of the desired phenomenon **(7)**. The actual positions of SNPs might not be helpful in some studies, especially in genome-wide scans. Moreover, according to the methodology used for a given study, the quality of each SNP might be very important. In a methodological context "quality" encompasses the nature of the flanking sequences and the polymorphic site itself, which account for their compatibility with the selected or available assay methods or techniques. Consequently, for a two-dye allele-determination system the polymorphic site must be strictly biallelic **(12)**. Therefore, every experimental protocol using SNPs as markers would require a selection of SNPs satisfying criteria for number and suitability specific for each experimental format. The extremely large number of published SNPs does not guarantee fulfillment of this prerequisite, especially for the less well-off labs, which cannot embark into the practice of confirming and verifying publicly available SNPs **(1)**. Such thoughts prompted us to simulate an experimental need for a SNP selection for the purpose of examining the efficiency ratio by assessing our selection procedures in relation to the quality of information offered by the NCBI SNP database.

Materials and Methods

SNP selection

For the selection of SNPs, the public SNP database of NCBI was used, found in the following e-address: http://www.ncbi.nlm.nih.gov/SNP/get_html.cgi?whichHtml=./maplists/maplist-newmap. Each study group was a gene, of any possible size, with three suitable NPs positioned one in its approximate middle and one on or near each of its two ends. For the SNP selection we followed a series of criteria aimed to facilitate the design of PCR primers through software. We also took precautions in view of a possible allele calling system, supposing it might entail an internal primer for techniques such as Single Nucleotide Extension and Allele-Specific Amplification. As a precaution in case the first method is to be applied, the polymorphic site was selected so as not to be A/T and C/G , because these pairs cannot be determined by two-dye system of SNE.

The criteria specified are as follows:

i) 150 bp minimum distance from neighbouring SNPs

ii) biallelic sites determinable by a 2-dye allele-calling system (A/G, C/T, T/G, A/C)

iii) flanking sequences which:

a) do not contain microsatellite DNA in the design frame (150 bp in each flank)

b) are not extremely "AT" or "GC" rich

c) do not have two complementary same- residue triplets 4 bp or more apart in a frame of 25-bp at the one flank of the polymorphic site and in two such frames at the other

The manual implementation of the above criteria meant that they could not be as stringent and thorough as a computerized and truly optimized procedure would both permit and require.

Primer design

For the step of the primer design, the software developed in-house was used.
The selected loci were saved in a *.txt (ASCII Text) file and the flanking sequences cropped to a maximum length of 150 bases. SNPs with flanking sequences of 50 bp and less (standard submissions by some labs) had to be enriched by the "Show Sequence" function of the SNP database. After successive zoom ins' the "Show sequence" projects the actual sequence in a variety of lengths, according to user specifications. By selecting FASTA format we were able to pinpoint the polymorphic site and the flanking sequences through standard "Search" functions and pick a larger part of blanking sequence. The lowest length acceptable by the software was 80 bases per flanking sequence. In the .txt file there were only the flanking sequences in uppercase, the polymorphic site and the sequence name (locus ID) of 10 characters, numbers and letters. We used as locus ID the standardized nomenclature developed by our lab specifically for the selection of SNPs for multi-locus studies instead of the official RS numbers of the SNPs. Special care was taken for the flanking sequences to be in uppercase, without any gaps (but between flanking sequence and polymorphic site) and no other characters than the four letters denoting bases (A, G, C, T). The program rejects any other character, as "N". The alleles of the polymorphic site had to be in capitals and separated by slash "/", without intervening spaces (as is its format in the database). In this study the amplified product was restricted to a maximum of 150 bp and the primers to a length between 20 and 25 bp (all these parameters- "r", "F", "f" respectively- are user specified and can be altered).

A high number of pilot runs allowed us to establish the following set of parameters as the most stringent possible combination of conditions for successful primer design in all our 150 loci: Melting temperature range between 73^{o} and 98^{o}C (parameters "T", "t" respectively), maximum sequence identity between 2 primers of 70% of the shortest one (parameter "m"), 3' to 3' end primer-primer interaction limited to 4 matching 3'-most residues, or to 7 if here is one internal mismatch (parameters "P", "p" respectively) and 3'-end to any (non-3'- end) primer-primer interaction of 8 matching consecutive residues or up to 10 with one internal mismatch (parameters "E", "e" respectively).

As the program picks sequences in random and compares them, the output is not the optimum possible with the submitted loci and flanking sequences, but the optimum that could be achieved with the sequences picked at random at the beginning of the run. This means that repeated runs might provide drastically different results.
Thus we run the program 25 times, with identical conditions and input file. Of the output files only 4 contained selection sets for all 150 loci without any drift from the specified parameters. All other output files had some loci for which the designed primers did not fulfill the specified parameters. Of the four sets, one was selected by virtue of the least number of primers with biologically undesirable sequence parts.

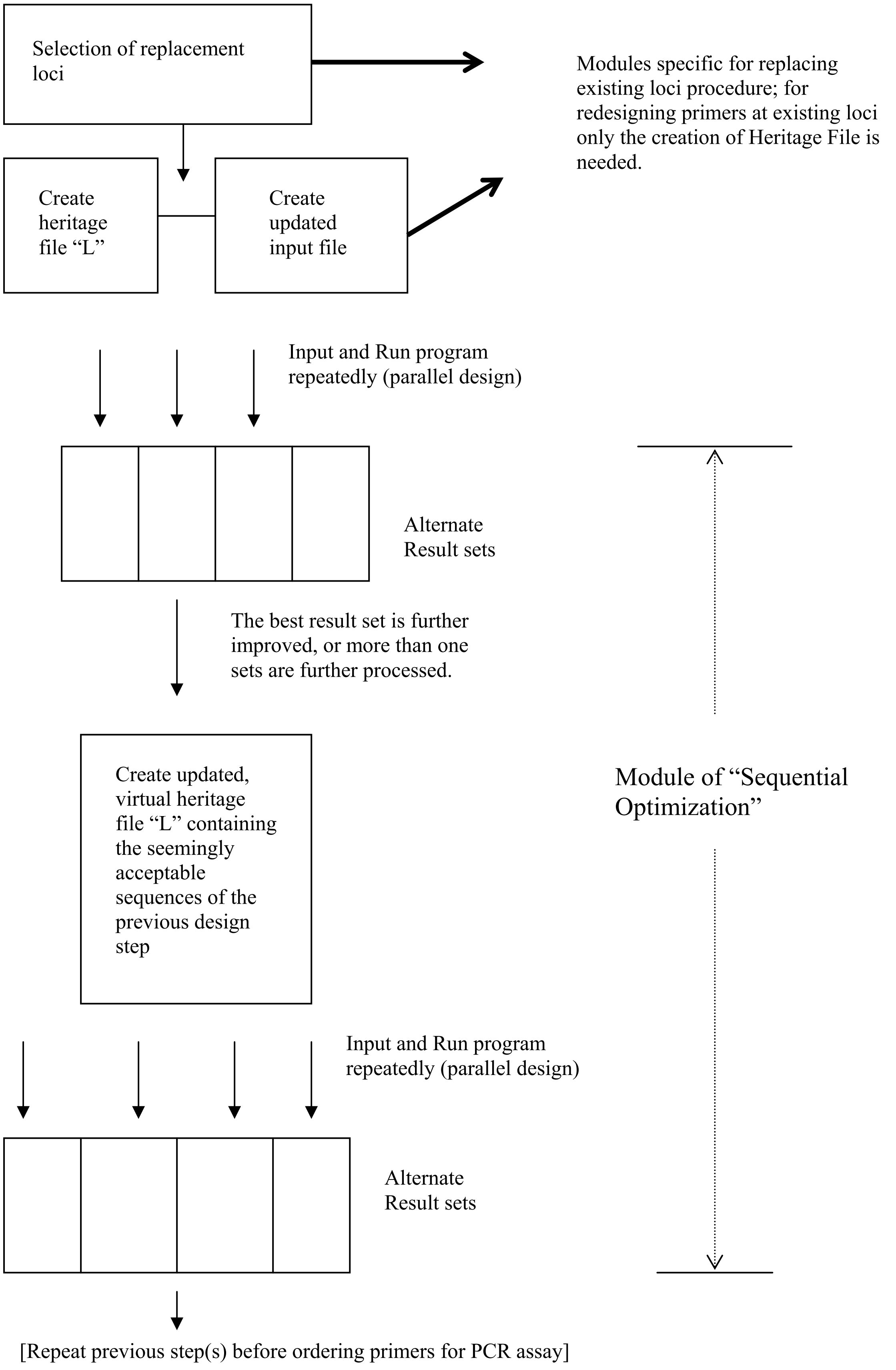
Primer editing

To redesign ("edit") the primers in some of the selected loci, (Design Cycle 2) the procedure was very similar to the original run for the primer design. The input file and the conditions were the same with the ones of the Design Cycle 1. The only exception was the inclusion of a new parameter ("L"), which permitted the edition function. This "heritage" parameter is a *.txt (ASCII text) file containing the codes and sequences of all the acceptable loci. When input into the software, all the contained primers will not be altered and new primers will be designed for the remaining loci (not included in the "heritage" file) so as to have the new primers in total computational compatibility with the existing ones. Since the procedure does not exclude anew selection of the existing primers for the editing loci, and in order to have the best possible primer pairs from a biological point of view, the program was run repeatedly with the same settings. This "parallel design" principle produced selection files with the new primers' sequences. Each of them (or alternatively, the most promising of them) could be further enhanced by the "sequential optimization" approach. According to the latter, from the newly designed primers the most promising primers are kept and incorporated into a new "virtual heritage" file. The defective ones are simply ignored, and the program is run anew with the new "virtual heritage" file as parameter "L". If there are still defective primers, there can be another step of the "sequential optimization" procedure, with ever-enriched "virtual heritage" files. Of course, in every step of the "sequential optimization" its is entirely possible (even advisable) to incorporate the "parallel design " principle, by running the program more than once with the same settings before selecting the output set which will be processed further by the next step of the "sequential optimization" procedure.

For the Design Cycle 4 (primers editing) the same procedure as with Design Cycle 2 was followed, but the input file was the one used in the Design Cycle 3 and the "heritage file" used also contained the acceptable primer pairs of the Design Cycle 3.
For Design Cycle 3 (replacement of SNP loci) the procedure of SNP selection was the same as described for Design Cycle 1. The primer design step was as in the Design Cycle 2, the only difference being that the input file contained the new loci instead of the ones that were rejected, and the heritage file was enriched with the acceptable primer pairs from the Design Cycle 2.

In Design Cycles 2 and 3 the "sequential optimization" was performed; in both cases only two such steps were used, whereas for the design cycle 4 we did not perform these steps, due to the minimal number of loci. In all Editing cycles (Design Cycles 2, 3, 4) the "parallel design" principle was used, with multiple runs of the program in each step. Though, we did not proceed to parallel processing; in each level the most promising output file was selected for further processing. Thus we simulated more accurately the needs and priorities of a true experiment, where cost and time are of prime consideration and decisions have to be made regarding which course to follow, as more promising.

Figure 1: Flow Chart of Primer Design (amendment) steps


Selection of replacement loci
Modules specific for replacing existing loci procedure; for redesigning primers at existing loci only the creation of Heritage File is needed.
Create heritage file "L"
Create updated input file
Input and Run program repeatedly (parallel design)
Alternate Result sets
The best result set is further improved, or more than one sets are further processed.
Module of "Sequential Optimization"
Create updated, virtual heritage file "L" containing the seemingly acceptable sequences of the previous design step
Input and Run program repeatedly (parallel design)
Alternate Result sets
[Repeat previous step(s) before ordering primers for PCR assay]

Primer handling and simplex PCR assay

For all 4 Design Cycles the steps after primer design were identical: From the "computationally perfect" output files, one was selected after visual browse for the lowest context of primers with biologically undesirable sequences. These primers were ordered at the 100 nanomole scale (IDT, Ia, USA) and were dissolved with 10 mM Tris-HCl (GIBCO BRL, Md, USA) buffer of pH 7.5, to a master solution of 100 μM. The buffer was autoclaved in 120^o C for 1h in liquid cycle. The volume for each primer was computed by an in-house software which is fed with the sequence and the manufacturer-provided OD value for each synthesized primer in a *.txt (ASCII text) file.

Working solutions were prepared for each locus, were the two primers of each primer pair were mixed together in 1:1 molecular ratio. An aliquot of the mix, typically 50 μl, was diluted by addition of four volumes of dd H_2O (i.e. typically 200 μl) and brought to a final concentration of 10 μM each (20 μM total primer concentration) in 2 mM TrisHCl. Simplex PCR was performed at a "T3 Thermocycler" (Biometra, Gottingen, Germany). The program used for amplification consisted of a preheat step at 94°C for 15min, 35 cycles of 94°C for 40 sec, 55°C for 1 min a ramping step of 0.2°C/sec from 55°C to 70°C and a final extension step of 72°C for 3min.

The 25-μl final volume reaction mixture contained 1ng DNA, 0.5 u HotStarTaq DNA polymerase (Qiagen, Ca, USA), and a final concentration of 100 μM isomolecular dNTP mix (GIBCO BRL, Md, USA), 100 μM Tris-HCl pH 8.3, 50mM KCl (J.T. Baker, NJ, USA), 10mM $MgCl_2$ (Sigma, Mo, USA), 0.1mg/ml gelatin (Difco, Mi, USA) 0.4 μM locus-specific two-primer mix. All reagents were added in a laminar flow hood by special pipette sets and autoclaved expendables; the target DNA was added last on benchtop.

Electrophoresis of PCR products and Visualization of gels

Standard 40% w/v Polyacrilamide/bis gels [1/1000 v/v TEMED (GIBCO BRL, Md, USA), 0.08% w/v Ammonium Persulfate (AMRESCO, Oh, USA), 19 :1 w/w polyacrilamide (Sigma, Mo, USA)/ bis (GIBCO BRL, Md, USA)] were run in 0.5X TBE (Boric Acid from Sigma , Mo, USA; EDTA from J.T. Baker, NJ, USA) for 25 min with 3.5μl of 33 ng/μl solution of pBR-322X *Msp* I (NEB, Ma, USA) as size marker. The standard well load was 6.5 μl of PCR product and 2.5 μl of dye [0.25% (w/v) Bromophenol Blue (Sigma, Mo, USA), 40% w/v sucrose (GIBCO BRL, Md, USA)].

The gels were stained in 0.5 μg/ml aqueous solution of EtBr (GIBCO BRL, Md, USA) for 5 min and visualized under UV at a GELDOC 1000 (Bio-Rad, Ca, USA) Gel Documentation System, using the manufacturer's "Quantity One" software.

When a PCR product appeared negative in the first, standard electrophoresis, a second one was performed with "double-load" (10 μl of product and 3.5 μl of dye) to discriminate between clear-cut negatives and primers displaying low efficiency (LE). Then, the locus was re-amplified, to exclude experimental error. A locus was deemed LE or Negative only if producing no product, or very faint product, in 3 successive simplex PCR assays.

In cases where the first electrophoresis revealed an extremely thick band of the expected length, a second run was performed with "half"–load (2.5μl of PCR product and the same volume of dd H_2O mixed with 2.5μl of dye). This permitted determination of multi-band products, with bands of similar but not identical sizes against clear-cut extra-high efficiency primer pairs.

Restriction Endonuclease Assay

When needed, REA digestions were carried out to determine wether the visualized band was the predicted one or a byproduct of similar size. For digestions we were using 10 μl of the PCR product and 1 μl (usually approx. 10 u) of the selected restriction endonuclease (NEB, Ma, USA) in a final volume of 15 μl, according to the specifications of the manufacturer; incubation was carried out in water-baths overnight. Enzyme selection had been carried out manually, by comparing the restriction sites of the enzymes already in the inventory through the suppliers' catalogues to the known sequences of the predicted PCR products. The digested product was run in standard polyacrilamide/bis gels as described above.

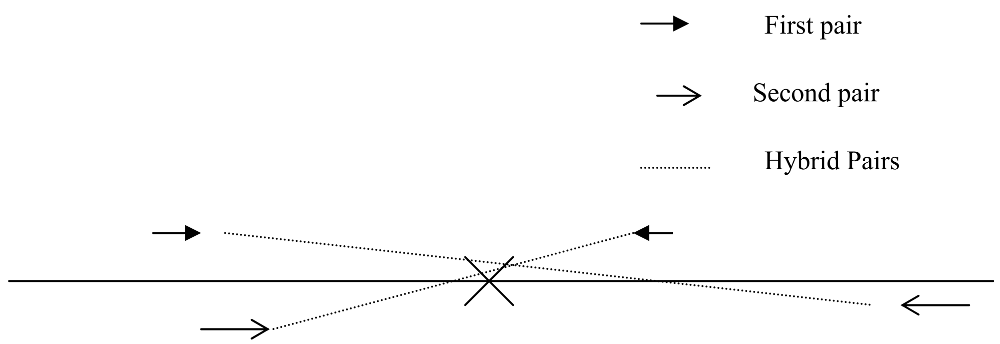
Virtual splicing

For the loci containing a part of their flanking sequence(s) unsuitable for PCR design (such as microsatellite, or other small tandem repeat sequences) and for countering the consistency phenomenon (exactly the same primer designed for a locus in multiple runs of the program), we reverted to the "virtual splicing method". The unsuitable, consistent or otherwise unwanted part is deleted of the locus sequence before submission to the primer design program. There can be more than one such "virtual spicing sites" in each of the flanking sequences of the locus. When the program designs the primers, it must be compared to both the virtual and the original sequence, to assure that they are not crossing the spicing sites, as this sequence is virtual and does not exist in reality. Each designed primer is acceptable only if it anneals between two virtual splicing sites. The amplified frame might well exceed the specified size limit (in our case 150 bp), by as much as the total length of the virtually spliced sequences found between the two primers.

Hybrid primer pairs

After a primer editing cycle (Design Cycles 2, 4) and if the product of the new primer pair has not been satisfactory, it is possible to use one primer from one pair and its countersense from another pair. The possibilities of obtaining an acceptable product per primer pair are thus multiplied. The amplified frame might be more than the specified size limit. Moreover, the primers of the hybrid pairs are not checked for interactions to each other and to other hybrid pairs.

Figure 2: The Hybrid Primer Pairs Principle



Hybrid primer pairs interaction elimination

The primers of the hybrid pairs that produce acceptable PCR products (in quality and size) are copied in another *.txt (ASCI text) file and fed in yet another in-house software, which shows annealing interactions between pairs of the submitted primers. The program projects pairs surpassing a threshold of successive matching residues in especially troublesome formats (3'-end primer sequence to 3'-end and 3'-end to any part of a primer's sequence) and the total number of matches between the two primers of the interacting pair. The operator accepts or rejects manually the primers according to criteria established for each study and with a view to eliminating as few primers as possible. In some cases eliminating one primer results in nullifying several interaction pairs and thus makes many primers acceptable. The eliminated primers have to be subsequently replaced.

Results

Our selection procedure allowed for complete coverage of all 50 needed study groups with 150 SNPs in total. The primer design software, after 25 runs, gave 4 selection sets with no computational flaws. From these, one was selected with a view to the fewest possible biologically flawed sequences. The products of the simplex assays were of a predicted size range between 89 and 149 kbp. The simplex PCR reaction assay resulted in 120 totally acceptable loci (sharp, single bands of the expected size) and 6 loci with multiple band products. The latter 6 cases, however, were deemed acceptable because the expected band was much more intense in Ethidium Bromide staining than the secondary products. So, the efficiency of the selection methodology was 84%, as a total of 126 assayed acceptable primer pairs out of 150 selected and designed. Of the 24 rejected loci (16% rejection rate), 3 primer pairs (2%) produced no bands or extremely faints bands in three simplex assays, 13 (8.66%) produced multi-band products, 5 (3.33%) produced a smearish or extremely fat band at approximately the expected size 2 (1.33 %) produced multi-band product with the main band being smearish. Low loads of the products with the smearish bands revealed more than one bands of similar but not identical size. Lastly, one locus (0.66 %) produced a multi-band product, which would have been acceptable, but the brightest band was clearly bigger than the estimated size, whereas there was no band of the estimated size. A search in the next build of the SNP database showed an ambiguity in the marking of the SNP's position.

The 24 unacceptable primer pairs were stemming from an equal number of SNPs positioned on 18 genes; of these, 6 genes housed two rejected SNPs, which meant that they should be eliminated as study units. The rest 12 genes each housed one rejected SNP.
The first editing cycle (second primer design cycle) redesigned primers for 23 of the 24 loci that failed in the original assays; for the 24th locus new primers could not be designed, as the software was consistently picking primers from exactly the same areas of the flanking sequences (maximum exhibited sequence difference 20% per primer in 20 runs of the program). This locus was tackled in due course by the "virtual splicing" concept. The new primer sets designed for the other 23 loci had a predicted product size range between 89 and 149 kbp. Of the 23 new primer pairs, only 8 (35%) cleared the simplex assay; 3 with sharp single band products and 5 with acceptable multi-band products. Only one primer pair gave no product (4.3 %) – incidentally, positioned on one of the 3 loci which were negative in the first cycle. Though, the total rejection rate of this cycle was 65%, more than four times the rejection rate of the first design cycle. After these results, we moved to the use of "hybrid primer pairs" for the remaining 15 loci. This effort allowed four more loci to be cleared through the simplex PCR assay to the collection.

Subsequently, we tried to replace the remaining 12 SNP loci (11 for which the new primer pairs had resulted in no improvement and the one for which no radically new primers could be designed). We were able to find suitable replacement loci, fulfilling all the original selection criteria for 9 of the 12 (75%).

The third design cycle was similar to the first and the rejection rate in the simplex assay was 22% (two unacceptable primer pairs out of nine, none with no visible product). One of the two initially unacceptable primer pairs was cleared though, after successful REA assay with *Dpn* II restriction endonuclease (final efficiency 88.8%, very close to the 84% of the First Design Cycle).

The fourth Design Cycle comprised the rejected locus of the third cycle and the 3 loci which did not even enter the 3rd cycle due to absence of suitable replacements. For one of them new primers could not be selected, as described earlier; for the other 3 new primer pairs were designed, but produced no acceptable products. The hybrid primer pairs method also failed to produce any clear-cut acceptable products, but one of the three loci was deemed acceptable after successful REA assay with the *Dde* I endonuclease. The other two loci were rejected and deleted from the collection, being positioned onto the same gene.

The locus for which no really new primer pairs could be designed (and no replacement SNP found, either) was tackled through the "virtual splicing" method; thus, after 4 runs an acceptable primer pair was designed by the software. The new pair produced a clear-cut acceptable product, but of much bigger size (approx 190 bp). When the original primer pair and the new primer pair were used in two hybrid primer pair reactions, one of them produced a single-band product of approx 170 bp, which was deemed acceptable. Thus, the end rejection rate was 2 SNPs out of 150 (end efficiency rate 98.66 %) and, since they were both positioned on a single gene, the study group deficit was only 1 out of 50 (end efficiency rate of 98%).

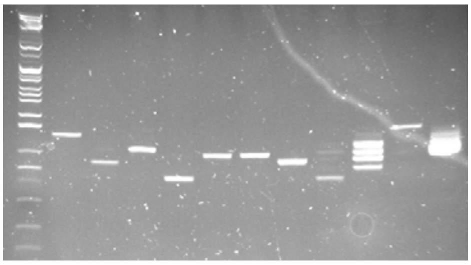


Figure 3: Typical PCR assay
Typical simplex PCR assay for 11 loci visualized in EtBr –stained polyacrilamide/bis gel. The first and third lanes from the right show multi-band products, making the primer pair unsuitable for incorporation into multiplex format.

Discussion

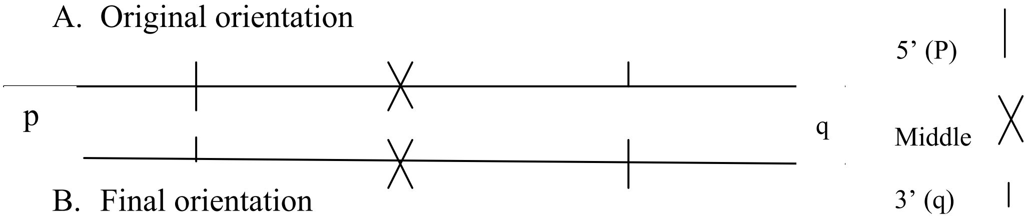
Two baseline requirements can be envisaged for SNP location: the first calls for SNPs on fixed positions within –part of- the genome, forming a physical map **(8)**. The second calls for SNPs located at relatively stable relative positions either to each other or to other genetic entities, such as genes and different kinds of genetic markers; such a set of SNPs and genetic entities can be defined as a "study group"**(4)**. It is obvious that the latter case would also greatly profit from absolute positional accuracy, but this is not a prerequisite, as happens in the former case. All that is needed is the detected and selected SNPs to remain at steady distances (or, even, at steady relative positions) to each other within the selected study groups.

Our efforts to simulate an experiment demanding high positional accuracy of the available SNPs came to abrupt end due to the dynamic (as yet) nature of the database. The required –and allegedly moderate- tolerance of a positional accuracy of half Mega base pair (Mbp) could by no means be secured between successive updates ("builds") of the database.

Thus we focused on the latter case. For our simulation we accepted the need of 150 SNPs selected so as to implement rather relaxed location criteria. As such, the location of three SNPs on a single gene was promoted as our study group because it is a realistic convention for a multitude of applications. The 3 SNPs of each gene were to be placed on or near each end and at the middle. Of course the location accuracy for each of these 3 sites was arbitrary and a function of the gene length rather than a more strict accuracy criterion. By accepting as ruling prerequisite the location of SNPs on or near a gene, we aimed at limiting the selection to rather well -mapped -in relative terms at least- SNPs, being anchored on rather well-chartered coding areas. The number of 3 SNPs per gene was promoted as it is the least number of points in a linear, vectored group needed to ensure the correct bearing of the vectored group in face of longtitudal (co-axial) positional uncertainty of high order[1].

[1] Single marker's positional uncertainty of absolute value possibly higher than the minimal distance of two successive markers in the selection group.

Figure 4: Vectoral (external) Shift Phenomenon



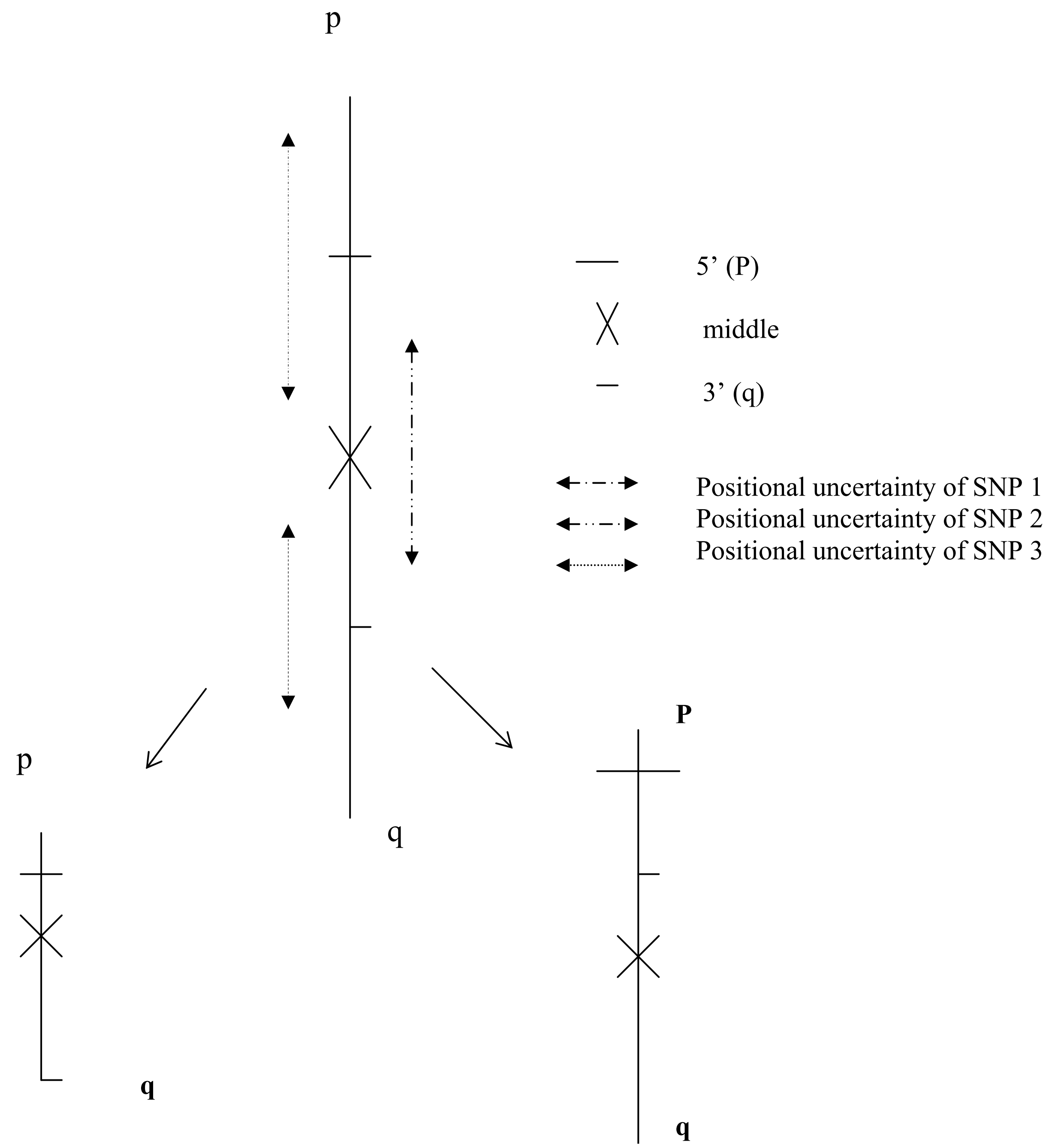
We also opted for all 150 SNPs (50 genes) to be syntenic due to the higher uniformity of the SNP quality on a gene compared to SNPs selected from throughout the genome (personal observations).

For most of the SNP-oriented methods an initial, locus-amplification step based on PCR is considered pivotal before the actual allele-determination step, for which there are quite some alternatives **(2), (5), (6)**. Thus, we decided that all the selected SNP loci should comply with a series of specifications aimed to facilitate PCR primer design. Since the way ahead for genome-wide scans seems to be multiplexed formats for the amplification steps, the primers were designed for use in such a format. This specification for "multiplex compatibility" stressed further the SNP selection step. For use in a multiplex protocol, the predicted PCR products should be more or less uniform in size, for a degree of uniformity in amplification efficiency; thus the software was set at a maximum product length of 149 base pairs, which was also specified as the minimal distance between a candidate SNP locus and its immediate neighboring SNP.

Finally, the methodology was judged for its success fraction in providing PCR primer pairs for the selected loci compatible to multiplexed formats, according to computerized criteria. The object of this study ultimately was to determine how many of the 150 needed SNP marker loci of specific qualities, could actually be brought to the actual experimental level. This fraction would provide an indication of the combined adequacy of SNP selection form the NCBI database and computerized primer design for multi-SNP-centered studies.

The acceptable primer pairs (and thus SNP loci) had to be cleared through a final Quality Control step, by being experimentally tested for actual amplification in simplex PCR reactions. A primer pair was deemed acceptable when the amplification assay (carried out with identical conditions for all the tested pairs) produced a sharp, clearly visible band of the predicted size with no or few secondary bands. In the case of one –or very few-secondary bands of comparable intensity, a Restriction Enzyme Analysis (REA) assay was carried out for the expected band. If only the expected band was digested, this was taken as an indication of low internal sequence homology and the primer pair was accepted. If even one of the secondary bands were digested, as well, this was considered an indication of higher internal sequence homology and the primer pair was rejected. All digestions were carried out with excessive amount of enzyme to exclude the possibility of incomplete digestions. Primer pairs with low efficiency (faint bands) or with high efficiency producing an extremely thick, or smearish band were rejected as well; the former as ineffective, especially in a competitive multiplex environment, the latter as indicative of multiple-loci amplification.

Figure 5: Longtitudal (co-axial) positional uncertainty of high order


p
5' (P)
middle
3' (q)
Positional uncertainty of SNP 1
Positional uncertainty of SNP 2
Positional uncertainty of SNP 3
q
p
q
P
q

Case 1:
Change of Distances

Case 2:
Change of relative positions (internal shift) on the vector

Of course, in this study the declared aim of picking SNPs from a decently "populated" chromosome and at the environs of the genes offered a rather unchallenging task. Things might be very different in other SNP selection formats for conceptually different studies. For example, a physical map of SNPs at pre-determined distances would face more adverse conditions of, and higher rigidity at, SNP selection; this is due to the local heterogeneity of the chromosomes (and the genome in general) as opposed to the distance-normality of a physical mapping system. In such a context the SNP selection will have to overcome special local difficulties such as the highly repetitive parts of the genome, areas with such extreme SNP densities, that PCR primer design is hindered and areas with extremely few or low-quality SNPs. Any systematic study for the projected adequacy of a SNP selection procedure, which must incorporate the PCR primer design, cannot be complete without such worst-case scenario data. The as yet dynamic and changing nature and the fluidity of the SNP database, though, negated our efforts to simulate and test this approach.

Even the favorable conditions of this study were not enough to secure an easy pick of SNPs and a trouble-free, one-step primer design procedure. Multiple actions were taken in a series of primer-optimization cycles to ensure the admittedly high accomplishment rates we show (49/50 in study units, i.e.98%, and 148/150 of selected SNPs, i.e. 98.66%). The concept of a multitude of small study units (in this case genes bearing 3 well-positioned and experimentally suitable SNPs) has undeniable merits but also serious drawbacks compared to a single, cohesive SNP selection concept.

Amongst the former, the lower susceptibility in positional uncertainty is of paramount importance and actually the main reason for the endorsement of this concept for our simulation experiment. Moreover, since the "study unit" is an entity and not a position, there's more flexibility in replacing it with another that fulfills the SNP-centered criteria. This flexibility offers a higher tolerance in specifications' deficit at the early stages of the procedure, and also offers higher robustness at the level of the study design.

On the other hand, prominent among the drawbacks is the "multiplier effect" in case of one SNP locus turning ultimately unusable. In an individual-SNP marker system, dropping a SNP means one marker less. Contrarily, in this system of interdependent markers, the impact of dropping one locus due to experimental inefficiency or incompatibility extends to the whole study group. Thus, the actual impact on the study may, *in extremis*, be multiplied by the number of loci per study group. For example, the worst-case scenario in this particular study was a net impact co-factor trice the actual one, since each study group contained three SNPs. The whole of the study group is compromised and in some cases must be discarded, which means a need for either replacement of the study group or acceptance of the shrinking of the whole initial collection. Alternatively, in some experimental designs, the "residual" study group can be used, but this is a straightforward degradation of both its informativeness for the study and of the quality and standardization of the collection as a whole.

The other, inherent drawback of the "study groups" concept is the very limited collateral coverage compared to the physical position marker system. Due to concept reasons as much as due to positional uncertainty, the study groups are tailor-made to provide necessary data for the entities of the specific study. This data is very unlikely to be of any usefulness for other studies. Contrarily, data amassed through a physical mapping system for a certain study can have numerous applications for different but similar studies, thus functioning as "usefulness multiplier".

The shift to the study groups meant –in actual experimental terms-that in the initial stage, that is at the selection of the SNPs, we had enough flexibility to attain a 100% selection implementation rate, whereas for other physical mapping systems a considerable deficit would have been observed even in this early stage, in the form of gaps in the coverage (personal observations), due to the lack of overlapping between successive "contigs" in the draft human genome sequence. This is an important problem whenever source data are concurrently subjected to reviewing.

{The computerized primer design program made possible the timely design of primers for a handsome number of loci and, the most important, with specifications permitting the multiplex formatting of all the separate PCR reactions in one tube. It is obvious that the level of complexity was exponentially increased by the latter specification, but multiplex formats emerge as the way to go for high throughput parallel experimental formats. Manual or computer-assisted design of these primers would be prohibitive in terms of time needed and sheer computational complexity. Of course the process could not be- and is not- fully automated. User interaction is demanded in the specification of the various settings. In this case the target is the best balance between stringent conditions and number of input loci. The strictest the conditions set, the less the flexibility for design of many, multiplexable, non-interacting primers. Another step the human intervention is needed is at checking the produced primer set. The program selects sequences and designs primers in order to match the interaction, Tm and maximum product length criteria, but not any criterion of biological usefulness. Thus it tends to include in the designed primers parts of long homologous runs or any other "weird" sequence phenomenon (ie tandem repeats). Such sequences, for a variety of reasons are undesirable for PCR primers, and it is for the human user to reject a bad primer or primer pair. The task gets more arduous, as these sequences are rather rare (having been eliminated in most cases in the first-manual- steps of SNP selection) and thus highly favored by the algorithm ensuring the designed primers' diversity.

The absence of an exclusion feature, coupled with the random picking procedure used by the program (the latter being a must for the timely development of the software and due to processing power limitations) made extremely difficult the avoidance of biologically unsuitable PCR primers. There were ways to overcome this problem, such as the repeated runs (both "parallel design" and "sequential optimization") and the "virtual spicing", but these –especially the latter- are extremely time-consuming and laborious and might be implemented in very few loci each time. If more than one-or a few- loci have to be treated likewise, the laboriousness of the task negates the advantages of the automated design to a considerable extend. On the other hand, the inclusion of an editing function at a time while the procedure was already on-track, permitted numerous amendments that would have been simply impossible without this feature. The editing function adds tremendous flexibility to the program's application, permitting less initial runs, since the biologically unacceptable primers can be rejected individually, in sequential optimization, without any need for re-running the program for all the loci, as was the case originally ("parallel design"). The "sequential optimization" procedure allows the phased and continuous improvement of the primer design and also permits the staged feeding of loci in the case of large collections. The latter allows a semi-sequential, "batch" processing, which is far easier and faster for given processing power than the submission of the whole collection from the onset. Pilot runs have shown that, if allowance is made in the stringency, this "batch input" allows primer design for up to 600 SNP loci simultaneously, whereas the "single input" practice had resulted in a maximum processing capacity of 300 SNP loci simultaneously -or even less, depending on the degree of compatibility among the loci of the collection. In more practical terms, without the editing function there would not be any possibility for re-designing primers, which, through the Simplex PCR assay were shown to be ill-suited to the task. This would have resulted in an efficiency ratio of 84% for SNPs and drastically less for study units (64%). These values of the first Design Cycle compare rather unfavorably to the 98%, which was the case for both figures after the exploitation of the editing feature and the four successive primer design improvement cycles. The editing/replacement function allowed the ultimate deficit to be trimmed to one-eight (1/8) of the original value (for the SNPs) and to one-eighteenth (1/18) for the study groups. It is also important to mention that the 24 primer pairs deemed unsuitable were on 18 genes (study groups). This meant not only that 18 study groups out of the original 50 were degraded, but also that 6 of them (one –third of the number, or 12% of the original collection) had ceased to exist as such since they bear 2 unacceptable SNPs out of a total of 3. Since it is the relative position of SNPs within the same study group that matters for such studies, and not each individual SNP, the third SNPs on these 6 genes were utterly useless. Thus, the 6 valid but useless SNPs should be counted on the total SNP deficit factor, which is brought to 30, or 20% of the initial 150, implying an adjusted efficiency ratio of 80% for SNPs.

The use of the Edit function for the Design Cycle 1 was impossible for it was not included in the original software version. But had it been included, it is dubious weather it would give many improvements, since it was the multi-target pairs (non-unique locations) which caused most of the problems, and not the-rather few-biologically unoptimized (in designed sequence) primers. Of the 36 primers of dubious sequence quality, only two belonged to troublesome pairs (one producing smearish band and the other producing multiple bands). Moreover, as it is, the efficiency percentage of the first Cycle is indicative (under conditions) of the **unadjusted experimental suitability ratio** in a SNP picking procedure; initial use of the Edit function would have degraded it to **adjusted experimental suitability ratio,** which is of very little indicative view as it is a function of the adjustment pursued by any operator individually and thus of no standard value.

After the first cycle of primer testing and since the editing function of the software became available, we had to decide weather to redesign primers at the existing troublesome loci or to discard and replace the loci proper for which no acceptable primers where designed. We initially opted for the first solution, as it is not always possible to replace a conveniently positioned and suitable SNP locus with another, fulfilling the same criteria. Moreover, a change in locus would demand differentiation of the design of locus-specific infrastructure for the allele determination step (whichever method might be used); this redesigned infrastructure must then be cleared through a complicated process to allow a high level of multiplexing, without any guarantees that it will be so succesfully. On the other hand, once the second pair of designed primers failed to produce acceptable results, and the alternate use of one primer from each of the two separate primer pairs (“hybrid primer pairs” concept) had also failed, it was thought that the only possible solution would be the actual replacement of loci. The latter offered theoretically much better probabilities of success (provided a suitable SNP locus could be found in the whereabouts of the discarded locus) against simple primer replacement and redesign; this is so because once a primer pair shows multiple products, it is very probable that a longer portion of the site (if not its entirety) has extensive homology with other parts of the genome. Our actual findings confirm the above hypothesis by showing failure rates of 16% and 11.8% (first and third primer Design Cycles respectively) for new loci against 65% for redesigning primers for the existing loci (the second Primer Design Cycle).

A slight problem occurred in the “hybrid primer pairs” concept. In each “hybrid pair”, each primer was computer-designed but not the pair. Although these primers were cleared, through software, form unwanted interactions with all other primers of their design batch, they were not cleared between themselves and other primer pairs of similar nature. Moreover, the expected product could be over the 149-bp size that the software permitted, as these pairs were not selected, neither cleared, as pairs by the software. For the latter problem we had no answer and accepted the ensuing difference of size, provided it was not more than 20 bp in excess. Longer products were deemed unsuitable for multiplexed PCR formats and the pairs rejected. This was due to an observed tendency for smaller sequences to amplify more efficiently in multiplexed, highly competitive environments.

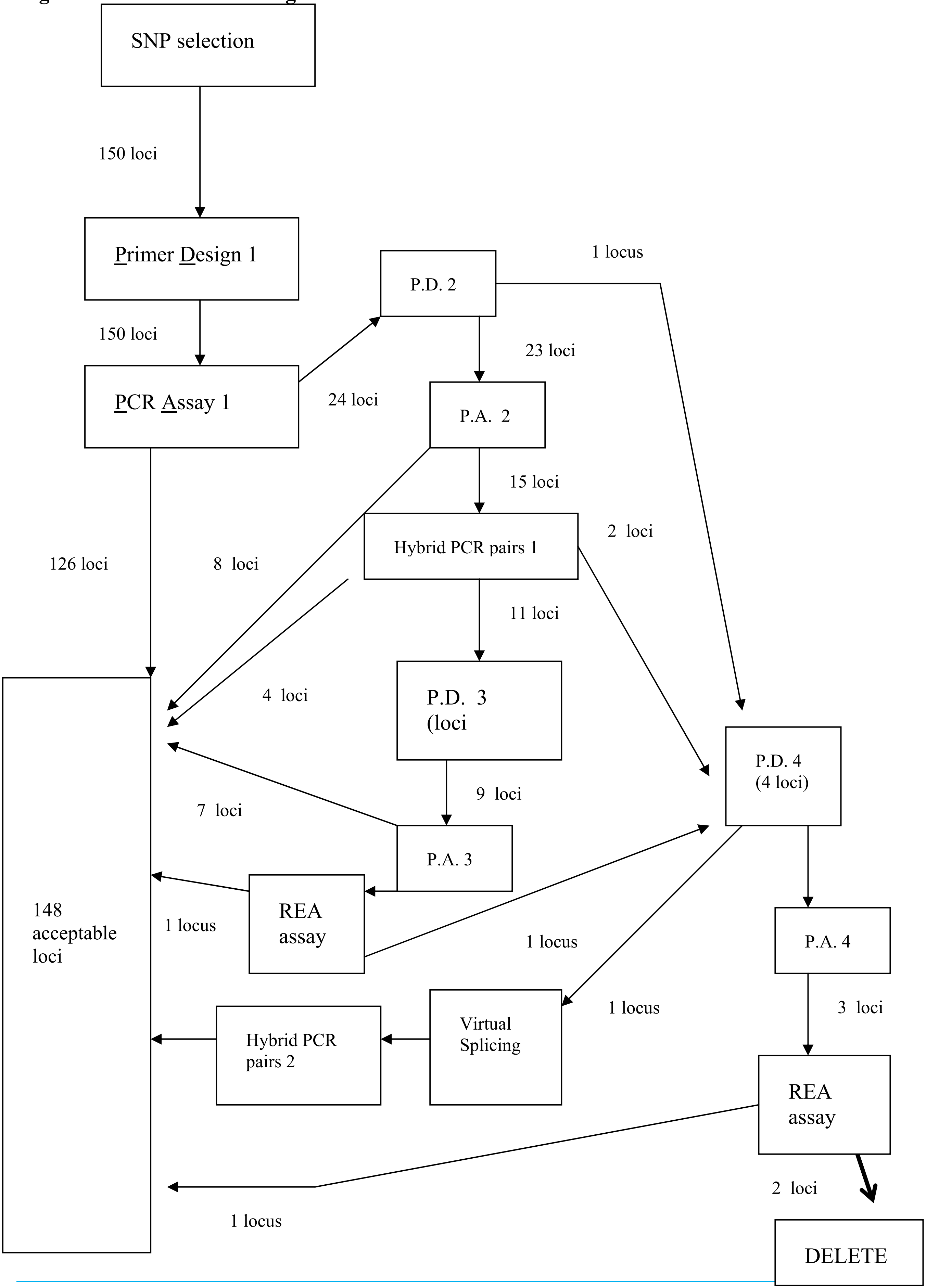
For the former problem, we tested all the primers of the successful hybrid pairs by eye for high homology and through another in-house software for interactions among themselves. This software checks for matching interactions on the submitted primers. These procedures were easily undertaken because we had to cope with a handful of loci, given that we confirmed the compatibility after a hybrid pair was proven acceptable. However, in cases where higher numbers of hybrid pairs must be processed there could be serious complications in the aforementioned procedure.}

The repetitive sequences, from what is mentioned above, were proved to be a major problem. The simplex PCR assay revealed an insignificant portion of “negative” results (no bands or extremely low efficiency), only 2% in the initial cycle. But, contrary to this, 21 loci (14% of the total) were deemed unacceptable due to multiple products. For multiple-band PCR product there’s always a high possibility that the homology will prove to be more or less external (i.e. at the ends, where the primers anneal) and in some cases it might be a case of “mirror priming”, (a key feature in RAPD formats) where one of the 2 primers of the pair amplifies by itself some sequences. The REA assays were exactly used to provide an indication of possible homology or differentiation internally, i.e. in the part between the primer target sequences.

Though, this approach, with all its merits and drawbacks, is of use only in the case of homologous PCR products of dissimilar sizes. In cases of actually repetitive sequences, the probability is that the PCR product will be identical, at times even to the polymorphic site. The SNP database charters only SNPs with flanking sequences allowing one or two perfect matches throughout the genome; the latter are clearly marked in all the database formats. But this procedure has resulted in characterizing just the SNP as a double –locus one, and not the whole site. We found out that in many cases the whole polymorphic site is not unique throughout the genome; though, since the polymorphic site occurs in only one of the replicate sites where this sequence is encountered, the SNP is marked in the database as unique. In the strict sense it is so; but the flanking sequences are not, and it mainly is through these flanking sequences that one can detect and process the polymorphic site.

The situation is further complicated by the vast differences of the submitted SNPs’ flanking sequences, which would naturally result in attesting different reliability levels during placing the discovered SNPs on the genome by the members of TSC. After our experience with this study, we introduced, as the only possible solution, a laborious but meticulous step of checking with the BLAST routine of NCBI the selected loci. Indeed, in some cases we used more than one databases and related functions and software options, to ensure the uniqueness of each selected locus within the human genome.

Figure 6: Flaw Chart of Integrated Procedure



The present study was conceived to allow for a practical, accurate and countable record of the expected efficiency in the incorporation of numerous SNP loci in the context of a pre-conceived study, without the vast resources of populous and expensively equipped labs. The hardware used and the methodologies were highly traditional and available everywhere. As this study was both a simulation, more or less, and a developmental study, in many cases we introduced new elements and improvements as it was already going on, such as the editing function of the primer design program and the REA assays for some multi-band products. We considered our objective completed at the point where primer pairs of shown acceptable levels of efficiency in simplex assays, were at hand, which were, nonetheless, designed for multiplex assays. The study of the efficiency of further steps, starting from the multiplex product and reaching the actual genotyping is altogether a very different issue, as it implicates diverse techniques with different efficiency characteristics and, in more cases than not, it will necessitate to the acquisition of new hardware, in some cases of high tech and high cost. Such considerations have been tackled successfully in other studies **(12)**.

On top of that, it is also a matter of which of quite a few procedures and methodologies one would select. The comparative tackling of such a diverse host of methodologies is an altogether special issue and formed no part of this study. Moreover, it is our immediate aim to assess the effect the aforementioned improvements (editing of primer design, BLAST and other alignment comparison steps) are going to have in the efficiency of SNP selection if integrated to the described procedure from the beginning of a study. As it is, although the original views might have been a bit optimistic, especially when one considers the levels of effort required for a decent SNP collection through the publicly available information sources, it is clear from the above that the use of large SNP collection is practical and in the realm of many mid-sized labs.

References

1. Ardlie K, Liu-Cordero SN et al (2001) «Lower-Than-Expected Linkage Disequilibrium between Tightly Linked Markers in Humans Suggests a Role for Gene Conversion» *Am J Hum Genet 69(3):* 582–589.

2. Dean FB, Hosono S et al (2002) "Comprehensive human genome amplification using multiple displacement amplification" *PNAS 92: 8* 5261-6,

3. 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;

4. Greenwood TA Alexander M et al (2001) "Evidence of linkage disequilibrium between the dopamine transporter and bipolar disorder" *Am J Med Genetics 105,* 145-51.

5. Huber M, Losert D et al (2001) "Detection of Single Base alterations in genomic DNA by Solid-phase PCR on oligonucleotide microarrays" *Analytical Biochemistry 299,* 24-30.

6. Hirschhorn JN, Sklar P et al (2000) "SBE-TAGS: An array-based method for efficient SNP genotyping" PNAS 97: 22, 12164-9.

7. Lai et al (1998) "A 4Mb high density SNP-based map around human APOE", Genomics 54;31-8.

8. Lehnert V, Holzwart J et al (2001) "A semi-automated system for analysis and storage of SNPs" *Human Mutation 17:* 243-54.

9. McCarthy JJ & Hilfiker R (2000) "The use of SNP maps in pharmacogenomics". *Nature Biotechnology 18:* 505-8.
10. Stephens JC Schneider JA et al, (2001) "Haplotype variation and Linkage disequilibrium in 313 human genes". www.sciencexpress.org/12 Jul 2001; 1-5
11. Venter JC. Adams, MD, et al (2001) "The sequence of the human genome". *Science 16, 291*: 1304-51.
12. 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

Acknowledgments
The author wishes to thank Drs Xiangfeng Cui, Sreemanta Pramanik and Hui-Yun Wang for their active and consulting help with the procedures, Drs Sioban McCarthy and Aristea Velegraki for their help with the linguistic aspects of the manuscript and Dr Honghua Li, under whose patronage, guidance and support this project was executed. This work was supported by grants from the National Human Genome Research Institute (R01 HG02094), and the National Cancer Institute (R01 CA77363 and R33 CA 96309), both to Honghua Li.

Σταδιακές διαδικασίες σχεδιασμού, σώρευσης και ποιοτικού ελέγχου Ολιγονουκλεοτιδίων για πολλαπλή γονοτυποποίηση Μονονουκλεοτιδικών Πολυμορφισμών με πολυπλεκτική PCR και μικροσυστοιχία ανάγνωσης Επιμήκυνσεων Μονήρους Νουκλεοτιδίου

Δρ 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 αξία των μονονουκλεοτιδικών πολυμορφισμών (SNP) ως δεικτών σε διάφορες μελέτες δεν εξαρτάται μόνο από τον απόλυτο αριθμό τους, αλλά και από το πολυπλοκότερο κριτήριο της διαθεσιμότητάς τους. Για κάθε δεδομένη μελέτη πρέπει να υπάρχουν αρκετοί SNP σε κατάλληλες θέσεις και ικανοποιώντες συγκεκριμένα γενικά κριτήρια καταλληλότητας. Η εφικτότητα της μελέτης εξαρτάται από την δυνατότητα αντικατάστασης των γενετικών δεικτών ώστε ο αριθμός τους να παραμένει σταθερός σε κάθε στάδιο της ακολουθούμενης πειραματικής διαδικασίας κατά την ανάπτυξη της συνολικής μεθοδολογίας. Επιχειρήσαμε μια πειραματική εξομοίωση, όπου καθορίστηκε αριθμός κριτηρίων για τον αριθμό και τις επιθυμητές ιδιότητες SNP που απαιτούνταν για μια υποθετική μελέτη και προχωρήσαμε στην επιλογή SNP και στη σχεδίαση εναρκτών δια λογισμικού για κάθε έναν εξ' αυτών. Ο καθοριστικός παράγων ήταν η καταλληλότητα των εναρκτών για πολυπλεκτική PCR. Τα διορθωτικά βήματα της ακολουθούμενης διαδικασίας επιλογής και οι ενημερώσεις του λογισμικού επέφεραν αποτελεσματικότητα των σχεδιαζόμενων εναρκτών της τάξης του 98,6%, καθώς 148 ζεύγη εναρκτών σχεδιάστηκαν ικανοποιώντας τις απαιτήσεις για πολυπλεκτικές PCR και ελέγχθηκαν επιτυχώς σε μονοπλεκτικές αντιδράσεις ποιοτικού ελέγχου, έναντι αρχικής απαίτησης για 150.

Λέξεις-κλειδιά

σχεδιασμός ολιγονουκλεοτιδίων, πολυπλεκτική PCR, επιμήκυνση μονήρους νουκλεοτιδίου, μικροσυστοιχία DNA, γονιδιακοί τόποι ανθρώπου, διαδικασίες ποιοτικού ελέγχου