

Abstract 1

A new online mutation repository for UK diagnostic laboratories

¹Gokhale, DA, ^{1,2}Devereau, AD and ³Taylor, GR.

1. National Genetics Reference Laboratory (Manchester), Regional Genetics Service, St. Mary's Hospital, Hathersage Road, Manchester, M13 0JH.
2. Department of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester, M13 9PL.
3. Yorkshire Regional Genetics Service, Ashley Wing, St. James's University Hospital, Beckett Street, Leeds.

Most NHS molecular genetics laboratories make regular use of online mutation databases to aid in the interpretation of mutation data prior to reporting. However, few laboratories have the time to submit details of the mutations that they regularly detect in the course of their work to these databases. This is despite the fact that the mutation data they collect is (i) likely to be of high quality (ii) has been subject to robust internal quality assessment and (iii) is of high relevance to others working in the diagnostic community.

We have now formulated a new database model for a diagnostic mutation repository based on the previous work carried out on the existing diagnostic mutation database (<http://dmudb.org>) and new collaborative work between the Leeds and Manchester laboratories. The above model has now been tested and forms the basis of a prototype web-based repository due to become operational in February 2004.

This repository has been designed to promote the collection of mutation data from accredited UK diagnostic labs and has the option of passing the information on to the HGVS repository for publication.

The process by which this development has taken place will be reviewed and the prototype web-based version of the mutation repository demonstrated.

Abstract 2

Reassessment of the p53 mutation database in human disease by data mining with a library of p53 missense mutations

Thierry Soussi¹, Dalil Hamroun², Shunsuke Kato³, Mireille Claustres² Chikashi Ishioka³ and Christophe Bérout²

1 Laboratoire de Génotoxicologie des tumeurs, EA3493 IC-UPMC, Hôpital tenon, Dpt Pneumologie, 26 rue d'Ulm, 75005 Paris, France thierry.soussi@curie.fr

2 Laboratoire de Génétique Moléculaire et Chromosomique, Institut Universitaire de Recherche Clinique, 641, avenue du Doyen Gaston GIRAUD, 34093 MONTPELLIER Cedex 5

3 Department of Clinical Oncology, Institute of Development, Aging, and Cancer, Tohoku University, Sendai 980-8575, Japan

p53 alteration is the most frequent genetic alteration found in human cancers. To date, more than 15,000 tumors with p53 mutations have been published leading to the description of more than 1,500 different p53 mutants (www.p53.curie.fr). In a recent report, we addressed various questions about p53 mutants found in the p53 databases (Soussi & Bérout, 2003). Indeed, the frequency of these mutants is very variable. The commonest variants are 175 G->A, (R175H), reported 688 times, 248 G->A (R248Q) 548 times, and 273 G->A (R273H) 468 times. These common variants are clearly real p53 mutants. Biochemical and biological studies have demonstrated that they have impaired DNA binding activity leading to a protein that is inactive for transactivation. They have also lost their growth arrest or pro-apoptotic properties.

At the other end of the spectrum, 585 variants are only reported once, 265 variants twice and 156 variants three times. It should be noted that only 12 variants are found more than 100 times, 194 variants between 11 and 99 times and 1,240 variants less than 10 times. This last series of rare variants corresponds to 29% of all mutations of the database.

So far, little is known concerning the biological significance of these rare variants, as the majority of biological studies have focused on classical hot spot mutants. In order to gain a deeper knowledge about the significance of these variants, we have cross-checked each mutant of the UMD-TP53 mutation database for its activity derived from a library of 2,314 p53 mutants representing all possible amino acid substitutions caused by a point mutation. The transactivation activity of all these mutants was analyzed with respect to 8 transcription promoters (Kato et al., PNAS, (2003)). Although the most frequent p53 mutants sustain a clear loss of transactivation activity, more than 50% of the rare p53 mutants display a significant activity. Analysis in specific types of cancer or in normal skin patches demonstrates a similar distribution of p53 loss of activity with the exception of melanoma, in which the majority of p53 mutants display significant activity. Our data indicate that p53 mutants represent a highly heterogeneous population with a large diversity in terms of loss of transactivation activity that could account for the heterogeneous tumor phenotypes and the difficulty of clinical studies.

Beroud, C. and Soussi, T. (2003) The UMD-p53 database: new mutations and analysis tools. *Hum Mutat*, 21, 176-181.

Kato, S., Han, S.Y., Liu, W., Otsuka, K., Shibata, H., Kanamaru, R. and Ishioka, C. (2003) Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci U S A*, 100, 8424-8429.

Abstract 3

The Y chromosome STR haplotype reference database (YHRD) – an online repository of worldwide collected population samples

L. Roewer^a, P. Croucher^b, M. Krawczak^b, M. Nagy^a and S. Willuweit^a on behalf of the International Forensic Y Chromosome Research & User Group (www.ystr.org)

^aInstitute of Legal Medicine, Humboldt University Berlin, Germany

^bInstitute of Medical Informatics and Statistics, Schleswig-Holstein University Hospital, Kiel, Germany

The collaborative YHRD project has been initiated by the *International Forensic Y Chromosome Research & User Group* (www.ystr.org) in 1996. The concept of continuous submission, banking and online presentation of individual Y-chromosomes typed at least for a core set of 9 discriminative Y-STR loci in quality approved labs creates in a short time period a large repository of Y chromosomal haplotypes from worldwide distributed populations. With about 150 visits/day the database is frequently used for statistical calculations in forensic, kinship and genealogical testing as well as in anthropological and population genetics research.

In February 2003, the YHRD has collected 21.407 Y-STR defined haplotypes from 198 populations from Europe, Asia, the Americas, Africa and Oceania contributed by about 150 academic institutions, forensic labs and national organisations involved in DNA data basing. Actually, a further rapid expansion of the database is to be expected, since commercially available Y-STR diagnostic kits including allelic ladders and internal controls are at hand.

Apart from its function as a tool to assess the statistical weight of male to male matches in forensic genetics the YHRD provide large sets of quality assured and carefully sampled Y chromosomal haplotypes for in-depth phylogenetic analyses of male population differentiation at a worldwide scale. This allows to approach one important problem in forensic and kinship diagnostics, the definition of largely homogeneous “reference populations” used to calculate matching probabilities.

To improve the user-friendly presentation of the data and to allow fast searches and statistical extractions via the internet a completely new database system (DBS) has been developed. The modular structure of the databases allows to embed an unlimited number of tools to interpret the data collection. The content of the database can be easily adapted by use of a content management system (CMS) which allows rapid editing without knowledge of programming languages. The database is safeguarded by a firewall. The maximum response time to a search request drops to about one second.

A number of new features has been added, most remarkably the “Haplotype neighbour search”, presenting the frequencies of all 18 +/- 1 step allele “neighbours” of a searched 9-locus haplotype. The reasoning behind this feature is, that a male genealogy is not only defined by one but rather a group of haplotypes emerged in time from each other by stepwise mutations. Likewise a Bayesian approach to calculate haplotype frequencies has been programmed to retrieve frequencies of rare haplotypes by comparison with its close neighbours (“Haplotype surveying method”).

In conclusion: A large, scientifically approved and user-friendly population database for individual Y chromosomes is only a mouse click away. The online availability of the database makes statistical calculations on basis of haplotypic data feasible for all experts working in the field of forensic or kinship diagnostics. The YHRD has also proven to be a valuable resource for the many laymen who are interested in their paternal genealogies. The treatment of this large data set by AMOVA (Analysis of Molecular Variance) and other biostatistical methods reveals new insights in the history and demography of European and worldwide male populations. A network of scientists and forensic analysts guarantees the quality and the timely update of the population data.

Abstract 4

A Phenotype Dimension to HGVbase

*Professor Anthony J Brookes (Anthony.Brookes@cgb.ki.se)
Karolinska Institute, Sweden*

In partnership with Heikki Lehtväslaiho et al (EBI, UK), my lab has established the Human Genome Variation Database (HGVbase) - a curated and annotated list of SNPs and other reported human sequence variants. Our operational policy has been to require evidence for the probable validity of variants before we include them in our resource. Variants and related annotations are harvested and submitted from a range of large and small discovery efforts, including primary literature.

Our intention is now to develop a phenotype dimension to HGVbase. To ensure the widespread acceptance and utility of the system we develop, we have organized PhenoFocus to create a consensus data model for phenotype data. This will then provide a robust basis upon which we can layer diagnostically relevant mutations, and the results of positive and negative association studies. Data submissions describing association studies will be passed forward to the journal Human Mutation (Wiley) to generate formal publications on behalf of the submitters.

We have so far established a basic model for phenotype data. This is founded on the use of EAV triplets (Entity, Attribute, Value) as a generic way to store diverse and potentially sparse descriptions of any one phenotype. We have defined a series of accompanying data fields to contextualize the EAV information and to define the attribute term to a practically useful degree. Most fields will accept only tightly controlled values, and use medically-related ontologies where possible. However, we will also allow a degree of flexibility for certain data elements as well as a free text field, so that the data model can evolve to suit the needs of real-world users. Our goal will be to develop an accompanying data submission tool that will carry extensive validation functions, so that data submissions can be made as convenient as possible. To date, we have built a first-version of the proposed phenotype system, and further modelled a 'MIGAS' data structure to elucidate a suitable 'Minimal Information for Genetic Association Study'.

Our experience suggests that data collection will require carrot and stick approaches, mediated in partnership with medical genetics journals and Grant Funding Agencies. To make this possible, we are involving the Human Genome Organization and the Human Genome Variation Society in the evolution of our plans and activities.

Abstract 5

Identification of 'forme frustes' and 'paucimorphisms' by population mutation scanning by meltMADGE: proof-of-principle using LDLR gene

1 N M Day, 1K K Alharbi, 1E Spanakis, 2L Haddad, 2R A Whittall, 1X Chen, 3H E Syddall, 3D I W Phillips, 5I Simpson, 2S E Humphries, 4G Davey Smith, 4D A Lawlor, 1S Ye, 3C Cooper, 4S Ebrahim

1Human Genetics Division, School of Medicine, University of Southampton, UK, Duthie Building (Mp808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK, 2Division of Cardiovascular Genetics, University College London, UK, Department of Medicine, British Heart Foundation Laboratories, Rayne Building, Royal Free and University College Medical School, 5 University St, London WC1E 6JJ, UK, 3MRC Environmental Epidemiology Unit, Southampton, UK, Southampton General Hospital, Southampton SO16 6YDUK, 4Department of Social Medicine, University of Bristol, UK, Canynge Hall, Whiteladies Rd, Bristol, BS8 2PR, UK, 5Wessex Cardiothoracic Centre, Southampton University Hospitals NHS Trust, UK, Southampton General Hospital, Southampton SO16 6YDUK

We describe here a new approach enabling identification of unknown mutations at the population level. We demonstrate its application to - 1. The definition of population-based 'reference ranges' for rarer sequence variation 2. The characterisation of 'paucimorphisms' (arbitrarily defined here as variants of rarer allele frequency, $0.05\% < q < 5\%$) 3. Research of contribution of 'formes frustes' milder mutations that are of significant relevance to the individual 4. The identification of severe mutations at the population level. The method, meltMADGE, is a reconfiguration of DGGE (using a thermal ramp rather than a urea gradient) enabling combination with microplate array diagonal gels (MADGE). Throughput per day per person is 4×10^4 96well gels in 2 2l tanks, representing 4,000 amplicons. Assays of LDLR exons 3 and 8 were validated in 460 familial hypercholesterolaemics with known mutations. We then applied the exon 3 assay in several DNA banks representing ~9,000 subjects with known cholesterol values and applied both assays in one DNA bank (n=3,600). In exon 3 we identified one known forme fruste mutation, P84S (n=1), also associated with moderate hypercholesterolaemia in this subject; an unknown silent variant, N76N (n=1); and known severe hypercholesterolemia splice mutation 313+1 G>A (n=2). Around exon 8 we identified a paucimorphism (n=35) at splice site 1061-8 T>C (known to be in complete linkage disequilibrium with T705I); and unknown splice 1186+11 G>A (n=1) and D335N G>A (n=1). D335N and a significant fraction of T705I subjects displayed cholesterol values above the 95th centile. Thus both severe, moderate and silent variants were identified. To our knowledge, this is the largest (or only) mutation survey to date of an unselected population sample. 5-20 million bases can be scanned per week at a running cost of 500Euros with initial hardware setup for 10,000Euros.

Abstract 6

THE MHC HAPLOTYPE PROJECT

COMPLETE MHC HAPLOTYPE SEQUENCING FOR SNP IDENTIFICATION: RESULTS FROM THE FIRST TWO HAPLOTYPES

Roger Horton

Wellcome Trust Sanger Institute, Genome Campus, Cambridge CB10 1SA, UK

on behalf of the MHC Haplotype Project.

The MHC Haplotype Project aims to sequence a 4.75 Mb section of human chromosome 6 including the major histocompatibility complex (MHC) using BAC libraries generated from eight homozygous/consanguineous cell lines carrying haplotypes selected for their autoimmune disease association. So far two cell line sequences have been completed and annotated to encompass all described splice variants of expressed genes and to define their complete variation content, revealing more than 18,000 variations (including >16,000 SNPs of which >340 were coding). SNP densities ranged from <1 to >60 SNP kb⁻¹. Variation analysis was conducted by comparing pairs of overlapping BAC sequences using the discrepancy list option of `cross_match`. The curated output, along with gene annotations, was read into an ACeDB database and dumped in gff (general feature format) for parsing into dbSNP submission files. Major indels were identified from breaks in the `cross_match` matches using RepeatMasker. Complete and accurate sequence data over polymorphic regions such as the MHC provide a definitive resource for the construction of informative genetic maps.

Web access: <http://www.sanger.ac.uk/HGP/Chr6/MHC/>

SNP submission: <http://www.ncbi.nlm.nih.gov/SNP/> under handle SI_MHC_SNP

Abstract 7

Application of the GENOLINK^(TM) Genotyping System in a Candidate Gene Association Study in Rheumatoide Arthritis

Holger Kirsten, Saskia Ruhland, Grit Wolfram, Peter Ahnert

Universität Leipzig, IKIT/BBZ, Johannisallee 30, 04103 Leipzig, Germany; ahnert@uni-leipzig.de

Rheumatoid Arthritis (RA) is a common complex autoimmune disease of unclear etiology. For several autoimmune diseases, studies suggest that the genetic make up of individuals plays a significant role in etiology and pathogenesis. For RA, HLA alleles have been most strongly implicated. However, they are estimated to account for only one third to one half of the total genetic risk. There is evidence that not a single gene or gene variant is responsible but genome wide variant patterns (GWVPs). GWVPs are certain patterns of specific variants of specific genes distributed throughout the whole genome.

The aim of our current research is to identify GWVPs associated with RA etiology and pathogenesis in a candidate gene association study. The selection of candidate genes is based on analysis of the current literature, both manually and aided by literature analysis tools. In the selection of polymorphisms, we aim to integrate current knowledge. The genotyping technology we use is the GENOLINK^(TM) Genotyping System by Bruker Daltonics. GENOLINK^(TM) comprises PCR, primer extension with special photocleavable primers, and MALDI-TOF analysis of extension products. To aid assay design and assay optimization we develop software tools, mostly based on Java^(TM). In our current study we investigate 78 different genetic variations, most of them SNPs, in 31 candidate genes in 450 individuals.

Our data show that GENOLINK^(TM) is a useful system for the analysis of genetic variation in the described setting.

Abstract 8

GENETIC VARIABILITY IN GENE SEGMENTS UPSTREAM OF CODING REGIONS

Labuda D, Langlois S, Gehl D, Beaulieu P, Lefebvre JF, Vasquez H, Moreau C, Labbé C, Theberge MC, Bourgoin S, Zotti C, Pastinen T, Lepage P, Hudson T, Dewar K, Sinnott D. Université de Montréal, Centre de Recherche Hôpital Sainte-Justine; McGill University and Genome Quebec Innovation Centre, Montreal PQ Canada

Diversity of human DNA is currently under intense scrutiny, leading us to better understanding of human origins, evolution, demographic history and population structure, essential to genetic epidemiological quest of complex diseases. For a dozen candidate genes, the *bona fide* regulatory regions arbitrarily defined as 2 kb segments directly upstream of their first codon were screened by dHPLC in 40 individuals of African, Middle-Eastern, European, East Asiatic and Amerindian descent. The polymorphisms were characterized by sequencing and subsequently genotyped in an extended panel of 80 individuals representing the same population groups. We found between 3 and 18 (average of 10) segregating sites in the 2 kb regulatory region. Nucleotide diversity estimates of 0.03 to 0.19 %, from both allele frequencies (average 0.08%) and the number of segregating sites (0.1%) fit well the genomic average. Based on the haplotype data for nine of the genes; their worldwide diversity was 0.65 (0.42 to 0.84) and their numbers from 4 to 21 (average 10) corroborate with the count of segregating sites, as if recombination played little role in diversifying these haplotypes. Yet, in three out of nine of these genes, recombinations (gene conversions) could have been documented suggesting that some of these *bona fide* regulatory regions recombine more often than the genomic average and may thus represent good candidates of recombination hot spots. The partitioning of haplotype diversity varies substantially among continental groups suggesting the possibility of diversifying and/or balancing selection acting on these genomic segments. Besides advancing our knowledge of the promoter regions, knowing these characteristics will find application in planning the association studies and in using the candidate gene approach in particular.

Supported by Genome Quebec/Canada and by Valorisation Research Quebec.

Abstract 9

MutRes and LsdbRes: extracting mutations to a central database

Pablo Marín-García¹, Albert J. Vilella², Anthony Brookes³, Heikki Lehtväslaiho¹

¹ *European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambs. CB10 1SD, United Kingdom*

² *Departament de Genètica Universitat de Barcelona Diagonal 645 – 08028 Catalunya – SPAIN*

³ *Karolinska Institute, Center for Genomics Research, Berzelius väg35, 171 77 Stockholm, Sweden*

MutRes and LsdbRes are two complementary projects that will provide an up-to-date relational database of Locus Specific Data Bases (LSDBs) and their content.

MutRes database, which compiles different mutation databases, has been rewritten and updated. Now, it contains nearly 300 up-to-date annotated LSDBs and related resources with detailed information. MutRes is a relational database with a Perl object layer providing web and command line interface to it. The description of the project, the code and the web interface for the database could be found at <http://www.ebi.ac.uk/mutations/mutres/>.

Besides the MutRes database, we are developing the LsdbRes database. LsdbRes is a collection of Perl Object-Oriented modules containing the parsers and modules to extract core mutation information from the LSDB sites listed in MutRes, as well as to create a standard annotate database with all the mutation for all the genes contained within this LSDBs.

The new MutRes and LsdbRes databases provide crucial metadata and tools for collecting and parsing LSDB mutation details and transfer them into HGVbase in near future.

We are now implementing a highly automated system to accomplish this process in clearly defined steps:

1. Maintain an up-to-date LSDB annotated database (MutRes);
2. Mirror LSDB data files;
3. Parse mutation entries from distribution files;
4. Keep track of changed entries since previous update;
5. Create standardized mutation representation for all the LSDBs genes, and categorize them;

And, in short term:

6. Validate mutations against reference data;
7. Write out new data for inclusion into HGVbase;

With this new pipeline, we will be ready to start distributing mutations for the benefit of a wider audience together with SNPs.

Abstract 10

Structural interpretation of mutations and SNPs in proteins using STRAP

Christoph Gille

Institute for Biochemistry of the Medical School Charite of the Humboldt University Berlin, 10117 Berlin, Monbijoustr. 2, Germany

Visualization of residue positions in protein alignments and mapping onto suitable structural models is usually the first step towards the interpretation of mutations or polymorphisms in terms of protein function, interaction and stability. For some proteins which are analyzed and screened for mutations three-dimensional structures are available. When the protein structures are not yet resolved structures of related proteins can frequently be found. A number of powerful protein viewers are available to the scientific community to view and render these structures. STRAP is a comfortable multiple sequence alignment editor for protein sequences. It can be started from <http://www.charite.de/bioinf/strap/>. Selecting and highlighting large numbers of residue positions in a protein structure can be time-consuming and tedious with the software currently available. Therefore we have designed several STRAP modules to address this issue. Now, STRAP can handle mutations on a nucleotide level as well as on an amino acid level and facilitates the import of mutation lists from databases. When the list contains nucleotide exchanges their position needs to be translated into the corresponding amino acid position by STRAP. For visualization STRAP employs external protein viewers. Currently, PYMOL, RASMOL and VMD are supported and others will be included in the future.

The analyses is guided by a wizard in STRAP. First, a close relative to the protein must be identified which has a known three-dimensional structure. A blast search against the collection of protein structures PDB is usually performed to find a suitable candidate. Second, both proteins must be aligned to transcribe the residue indices of the analyzed protein into residue indices of the protein with the known structure. Third, these residue indices need to be translated into residue numbers. This is because protein viewers usually do not understand residue indices counting from 1 to the number of amino acids. Instead, they address certain amino acids by using the residue numbers recorded in PDB-files. Finally the residues need to be highlighted in the structure either by choosing a different representation or by attaching text labels. Using Strap we analysed a mutation in the cardiac beta-myosin heavy chain Ser642Leu. It was found in a patient presenting with dilatative cardiomyopathy. Usually, mutations in this cardiac protein lead to cardiac hypertrophy. Using the new tools we were able to explain the unusual phenotype of the patient by the location of the muted residue Ser642Leu in the 3D-structure.

Abstract 11

Mapping SNPs and Locus Specific Mutations to Protein Sequence Structure Data

Antonio Cavallo and Andrew C.R. Martin, University College London

A project has been started using HGVbase data set as primary source for nucleotide mutation. The goal for such a project is to link single nucleotide polymorphisms to phenotype alterations, in particular to assess the structural mutations.

To achieve the result a data pipeline has been planned (and partially completed) bringing together information from mutation databases, gene annotated sequences sources and structural protein information. The steps involved are, in order:

- Get data for mutation (HGVbase)
- Retrieve the gene sequence where the mutation did occur (EMBL)
- Map the mutation to the protein structure (Swissprot and PDB)

Every step in this data flow has been checked in order to bring into the next step a validated set. Between stages, where appropriate, a control has been made to eliminate eventual "errors" related to data base version mismatch or update. Every single step has been as much as possible automated in order to accommodate update to the data bases.

Here we would like to present:

- The results about HGVbase data set validations are presented: the reports will discuss about the entries that are not entirely correct from the point of view of our analysis.
- The coherency problem between the reference inside the HGVbase data base and the EMBL: we show how we corrected for mutation positions not updated
- The mapping problem between an annotated sequence and the structural information (PDB).

References:

- 1 Fredman, D. et al., *HGVbase: A human sequence variation database emphasizing data quality and a broad spectrum of data sources*, *Nuc. Acids Res.*, 30:387-91 (2002).
- 2 Kulikova, T. et al., *The EMBL Nucleotide Sequence Database*, *Nuc. Acids Res.*, 32:Database issue **D27-D30** (2004).
- 3 Boeckmann B. et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003*, *Nuc. Acids Res.*, 31:365-370 (2003)
- 4 *Berman, H.M.* Et al., *The protein data bank*, *Nuc. Acids Res.*, 28:235-242 (2000).

Abstract 12

STATISTICAL PREDICTION OF PATHOGENIC VARIANT SITES IN HUMAN MITOCHONDRIAL GENOMES

Marcella Attimonelli⁽¹⁾, *Matteo Accetturo*⁽¹⁾, and *Daniela Lascaro*⁽¹⁾

⁽¹⁾*Dipartimento di Biochimica e Biologia Molecolare, Università degli Studi di Bari, Via Orabona, 4 – 70126 Bari, Italy, m.attimonelli@biologia.uniba.it*

Introduction

Mitochondrial DNA disorders – disorders associated with dysfunctions of the oxidative phosphorylation system (OXPHOS) – are caused by inborn metabolism errors and have an estimated frequency of 1 out of 10000 live births.

Due to the relevant role played by the OXPHOS system in ATP production, causes and effects of mitochondrial disorders are highly heterogeneous and complex [1]. Major origin of mitochondrial disorders is in both nuclear and mitochondrial DNA mutations. Although prenatal diagnosis is routine for nuclear DNA mutations, the cases of prenatal diagnosis of mtDNA mutations are rare, even though urgent, as no real therapies exist [2]. However thanks to bioinformatics support, the gap may be reduced in a short time. Indeed, up to now, the pathogenicity of mtDNA mutations has been, in most cases, prevalently validated by their segregation with the disease and their consequent loss of function when the mutation involves a structural gene, but no systematic statistical analysis of the mtDNA SNPs has been performed. Moreover the criteria commonly followed to associate a mutation to a given pathology are:

- aminoacidic change in a strictly conserved site;
- presence in patients only;
- heteroplasmy condition;
- presence in phenotypically similar, but ethnically different families.

However a strict correlation mutation-phenotype in patients is not always verified.

Here we propose a statistical approach aimed to contribute in the estimation of the pathogenic variation sites.

The analysis is based on the estimation of site-specific relative variability in a sets of homologous sequences, through the application of SiteVarProt [3] and SiteVariability [4] softwares, in order to infer a correlation between site variability and pathogenicity of a given mutation.

Methods

Site-specific variability indexes have been calculated starting from nucleotidic and aminoacidic multialignments, through the application of SiteVariability and SiteVarProt softwares respectively. Site-specific relative variability values for each i_{th} site (φ_i) of a dataset of N sequences, have been estimated according to the following formula:

$$\varphi_i = \frac{\sum_{j=1}^{N(N-1)/2} \delta_{ij} / K_j}{K_j}$$

where, as far as nucleotidic sequences are concerned, d is a parameter assuming value 1 when the variation is present and 0 elsewhere in the position i of the j pair sequences and K_j the mean genetic distance calculated for the j pair on the entire alignment with the GTR model [5] [6], while as far as aminoacidic sequences are concerned, d is a Blossum-like index (giving the level of similarity between two aminoacids) for the position i of the j pair sequences, and K_j the mean genetic distance calculated with the Kimura model. In both cases values of φ_i are normalized respect to the maximum value of variability (φ_{max}) calculated for that particular dataset of sequences, obtaining a new value φ_i in order to make site variability indexes comparable between two or more dataset of sequences.

$$\varphi_i = \varphi_i / \varphi_{max}$$

Sample

Two datasets have been used. The nucleotidic site variability estimate has been performed on mtDNA sequences of the 13 mitochondrial genes coding for OXPHOS proteins, belonging to 687 human subjects from different geographic origin. Most of the sequences have been retrieved from literature except for 7 belonging to West New Guinea individuals, which have been sequenced in our laboratory as part of a broader complete mitochondrial genome sequencing project. Whereas the aminoacidic site variability estimate has been calculated on the same set of sequences as far as human genomes are concerned, plus mtDNA sequences of the 13 mitochondrial genes coding

for OXPHOS proteins belonging to 60 mammalian different species retrieved from AMmtDB database [7].

Results

The comparative analysis of site variability patterns together with the association of MITOMAP [8] data relative to polymorphic and pathological mutations, has allowed to infer a relationship between the variability of a given position and the pathogenic potential of the corresponding mutation.

Table 1. Example of mutation classification on the basis of site variability data and their association with MITOMAP data.

FP=frequent polymorphism; PPP=Potential Pathological Polymorphism; RP = Rare Polymorphism; PM=pathological mutation;

MERRF = Myoclonous Epilepsy with Ragged Red Fibers.

| ATP8 | | | | | | | | |
|------------------|------------------|--------------------|--------------------|-------|------------------------|------------------------|----------------------------|-------------|
| aminoacidic site | nucleotidic site | nucleotidic change | aminoacidic change | index | human nucleotidic var. | human aminoacidic var. | mammalian aminoacidic var. | annotations |
| | 8414 | C-T | L-F | 0 | 0,61 | | | FP |
| | 8415 | T-C | L-D | -4 | 0 | | | |
| 17 | 8416 | | | | 0 | 1,000 | 0,601 | |
| | 8426 | T-C | F-L | 0 | 0 | | | RP |
| | 8427 | T-C | F-S | -2 | 0 | | | PPP |
| 21 | 8428 | T-C | | | 0,095 | 0,000 | 0,191 | RP |
| | 8429 | | | | 0,043 | | | |
| | 8430 | | | | 0 | | | PM |
| 22 | 8431 | C-T | S-L | -2 | 0 | 0,067 | 0,480 | (MERRF) |

This kind of approach has allowed us to define 4 different classes of mutations as shown in table 1, defined by their own site variability range.

Moreover nucleotidic and aminoacidic site variability patterns of the 13 mitochondrial coding for proteins genes have given the possibility to distinguish between high variable genes and low variable genes, in order to assess their proneness to accumulate pathogenic mutations, also in relationships with the previous defined classes of mutations.

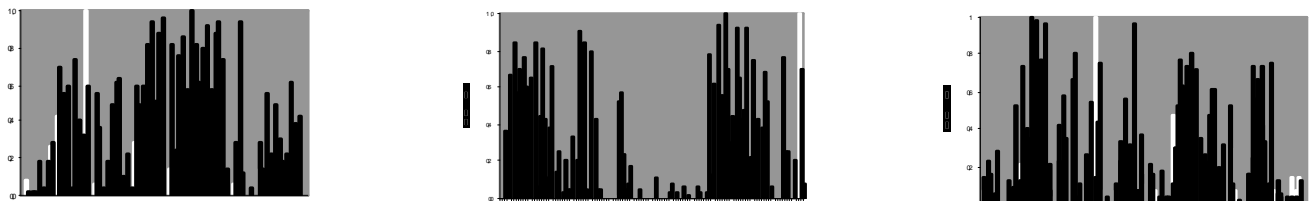


Fig.1. Aminoacidic site variability pattern of ATP8, ND3, COX3; in white human aminoacidic variability, in black mammalian aminoacidic variability.

We can conclude that this new approach could contribute to complete the knowledge about mitochondrial disorders, giving the possibility to shed light on the pathogenic potential of new mitochondrial mutations.

References

- [1] E.A. Schon, E. Bonilla, S. DiMauro, Mitochondrial DNA mutations and pathogenesis. *J Bioenerg Biomembr*, 29:131-149, 1997
- [2] J. Smeitink, L. van den Heuvel, S. DiMauro, The Genetics and Pathology of Oxidative Phosphorylation. *Nature Reviews Genetics*, 2: 342-352, 2001
- [3] D. S. Horner, G. Pesole, The estimation of relative site variability among aligned homologous protein sequences. *Bioinformatics* 19: 600–606, 2003
- [4] G. Pesole, C. Saccone, A novel method for estimating substitution rate variation among sites in a large dataset of homologous DNA sequences. *Genetics*, 157: 859–865, 2001
- [5] C. Lanave, G. Preparata, C. Saccone, G. Serio, A novel method for calculating evolutionary substitution rates. *J Mol Evol*, 20:86-93, 1984
- [6] C. Saccone, C. Lanave, G. Pesole, G. Preparata, Influence of base composition on quantitative estimates of gene evolution. *Methods Enzymol*, 183:570-583, 1990
- [7] C. Lanave, S. Liuni, F. Licciulli, M. Attimonelli, Update of AMmtDB: a database of multi-aligned metazoa mitochondrial DNA sequences. *Nucleic Acids Research*, 28:153-154, 2000
- [8] A.M. Kogelnik, M. T. Lott, M. D. Brown, S. B. Navathe, D. C. Wallace, MITOMAP: a human mitochondrial genome database—1998 update. *Nucleic Acids Res*, 26:112-115, 1998

Mutational Analysis of ATM Gene within Iranian Patients with Ataxia Telangiectasia

*H. Atashi Shirazi*¹, *B. Bayat*², *S. Ahmad Aleyasin*², *A. Farhoodi*³, *M. Moin*³, *M. Hossein Sanati*

1) Tehran Khatam University, 2) National Research Center for Genetic Engineering and Biotechnology (NRCGEB), 3) Children Medical Center

Ataxia telangiectasia is an autosomal recessive disease. More than 100 mutations in ATM gene have been reported which cause ataxia telangiectasia. Identifying common mutations in each population provides information for genetic counseling, prenatal testing and carrier detection. This study is to permit detection of ATM mutations in Iranian patients. We screened some Hotpoint exons and introns of five A-T patients and their families. Genomic DNA was extracted from their blood samples and the specific regions of the gene were amplified using PCR. Single strand conformation polymorphism (SSCP) was used for mutational analysis of amplified fragments from genomic DNA. We screened 11 terminus exons in which some were overlapped with introns. We identified polymorphic bands on SSCP gels after staining them with silver stain. Among 14 blood samples we studied, 5 were homozygotes for the polymorphism, 5 were heterozygotes, and 4 showed no polymorphism compared to negative control bands. In this study, polymorphism has been identified in exons 40, 41, 58, and 62. Homoallelic mutations would be expected to occur in consanguineous population, whereas heteroallelic (compound) mutation should be frequent among unrelated individuals.

KEY WORDS: Ataxia telangiectasia; ATM gene; Polymorphism; SSCP analysis

Abstract 14

Rapid genotyping of blood group antigens using multiplex PCR and DNA microarray

¹**Sigrid Beiboer**, ¹Tinka Wieringa-Jelsma, ²Petra Maaskant-van Wijk, ¹Ellen van der Schoot, ¹Dirk Roos, ³Johan den Dunnen, ¹Masja de Haas

¹Sanquin Research at CLB and Landsteiner Laboratory, Academic Medical Centre, University of Amsterdam, Plesmanlaan 125, 1066 CX Amsterdam, The Netherlands, ²Sanquin Blood Bank South West Region, Wytemaweg 10, 3015 CN Rotterdam, The Netherlands, Leiden Genome Technology Centre, LUMC, Wassenaarseweg 72, 2300 RA Leiden, The Netherlands

In the Netherlands, about 500,000 people volunteer to donate blood. Each year, 60,000 new donors come forward. For transfusion, blood of all these donors is currently serologically typed for only a few of the about 60 relevant blood group systems. Due to high costs and absence of test reagents only a subset of these donors is tested for more systems. Incomplete typing can lead to transfusion reactions. Therefore, it is our aim to develop a high-throughput technique to genotype by DNA microarray the whole donor cohort for 60 blood group systems. This means that 60,000 donors a year will have to be typed after two different donations, that is 138,000 genotypes each week.

The molecular basis for most blood group systems is known. Most blood group antigens are bi-allelic and are the result of a single nucleotide polymorphism (SNP). These SNPs are used for genotyping. After DNA isolation, gene fragments containing the SNP are amplified by PCR. To this end, a multiplex PCR has been developed to both amplify and fluorescently label gene fragments of 18 blood group systems in one reaction tube. The PCR products are then heat-denatured and hybridised to the DNA array without further purification. On each glass slide 12 arrays are present, containing spots of short (17-29 nt) allele-specific oligonucleotides. For each blood group 20 different oligonucleotides are spotted: 10 for each allele, sense and antisense. The allele-specific oligo hybridisation method (ASO) is used to discriminate between the two blood group alleles. A blind panel of 58 donor samples has been genotyped for 6 different blood group systems. For only one sample a discrepancy was found in one blood group, which urged us to adjust the scoring criteria and to validate the new format. These results show that the microarray will provide a reliable and fast procedure, which can be further improved to obtain the necessary throughput. The availability of a completely genotyped donor cohort will facilitate the selection of correct donor blood and improve the safety of blood transfusion.

Abstract 15

Expression cloning of T4 endonuclease VII and development of a MADGE-based heteroduplex cleavage protocol for economical high throughput mutation scanning

1,2[M J Smith](#), 1G Pante-De-Sousa, 1X Chen, 1I N M Day, 1K R Fox

1Human Genetics Division, School of Medicine, University of Southampton, UK, Duthie Building (Mp808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK, 2Division of Biochemistry & Molecular Biology, School of Biological Sciences, University of Southampton, UK, Bassett Crescent East, Southampton SO16 7PX UK

Mutation scanning is an important objective both in research and diagnostics. Identification of unknown mutations has been generally laborious and costly, whether by direct sequencing or scanning techniques. The cost is usually tens or hundreds of fold greater than the cost of PCR generation of the amplicon. T4 endonuclease VII is known to cleave heteroduplexes on one strand. Gel electrophoresis of the denatured (end-labelled) product both identifies presence of heteroduplex in the amplicon; and from the size of the cleaved single strand gives an estimate of the location of the heteroduplex mismatch. We have combined denaturing microplate array diagonal gel electrophoresis (MADGE), fluorescent end labelling of amplicon and endoVII cleavage to create a protocol which will be capable of examination of four thousand amplicons per day per worker. The T4 endoVII gene has been recloned and expressed in a His-tag vector in order to achieve convenient high yield production of the cleavage reagent. The endoVII-MADGE method has been validated by comparisons with endoVII-cleavage and capillary electrophoresis using a set of known SNPs and mutations previously validated both by direct tests or SSCP, meltMADGE and direct sequencing. It has also been validated by repeats of all protocols by independent workers. Hardware setup cost is approximately 10,000Euros and running costs around 500Euros per week, including PCR and electrophoresis. This protocol complements meltMADGE as an approach incurring little more than PCR costs and opening the way both for extensive studies of unknown and rare mutations in large population (as well as diagnostic) samples and for cost efficient high throughput which will generate population 'reference range' information important in interpreting the spectrum of sequence change such as aminoacid substitutions encountered in diagnostic practice.

Abstract 16

Definition of population 'reference range' for sequence diversity of MC4R gene using meltMADGE: two 'paucimorphisms,' occasional 'private' mutations and anthropometric consequences

1K K Alharbi, 1E Spanakis, 1S D O'Dell, 2A Aihie Sayer, 2C Cooper, 2D I W Phillips, 1I N M Day

1Human Genetics Division, School of Medicine, University of Southampton, UK, H Duthie Building (Mp808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK, 2MRC Environmental Epidemiology Unit, Southampton, UK, University of Southampton, Southampton General Hospital, Southampton SO16 6YD, UK

Identification of mutations has remained laborious and expensive and only viable for small numbers of subjects or cases. Population based 'reference ranges' of rarer sequence diversity have not been available. However, the research and diagnostic interpretation of sequence variants can be crucially dependent on such information. We have developed a high-throughput system, meltMADGE, which reduces scanning cost to a fraction of PCR cost (1/7) rather than a multiple of it (10-100x). MeltMADGE combines the properties of Microplate Array Diagonally compatible PAGE gels (Gaunt et al, NAR 2003, 31 e48-10) with a reconfiguration of denaturing gradient gel electrophoresis, such that the independent (denaturing) variable is a DNA melting thermal ramp in time instead of a chemical (urea) gradient in space. The temporal dimension of the melt then permits use of high density 2D arrays of electrophoresis tracks, such as used in MADGE. Two heteroduplex and two homoduplex bands should resolve from a heterozygote amplicon. Throughput per worker per day is 40x96well gels=4,000amplicons, using two 2l tanks taking 10gels each for 2hr runs. We developed six assays representing the MC4R gene and examined a population sample of 1,100 subjects. Two 'paucimorphisms' were identified (V103I in 27 subjects and -178A>C in 30 subjects). Anthropometric studies of these variants have the power to detect, for example, BMI effects as little as 0.5units. Two rare variants were also identified, one previously described (T112M), one unknown (A87D) ? BMI of 31.5 in the latter might point to mild functional effect, although high birthweight (4763g) argues against postnatal hyperphagia. Approximately 3million bases were scanned in a total time of 1week at a total cost around 500Euros. Expansion to much larger survey sizes would be straightforward.