2

Family Based Studies in Complex Disorders: The Use of Bioinformatics Software for Data Analysis in Studies on Osteoporosis

Christopher Vidal and Angela Xuereb Anastasi
University of Malta,
Malta

1. Introduction

Complex diseases are common within human populations and communities and pose a great burden not only to affected individuals, but also to society and the health system. Disorders such as chronic heart disease, diabetes, Alzheimer's, epilepsy and many others, are caused by complex interactions of a number of genetic and environmental factors. This makes the identification of the responsible genes difficult if using the same methodologies used for monogenic diseases. For more than fifteen years there has been a collective effort by researchers from around the world to identify genes and genetic variations that increase the risk for osteoporosis and fractures in ageing populations to identify novel therapeutic and prognostic targets, but predominantly most studies have been inconclusive. Genetic heterogeneity between different populations is the main factor responsible for this lack of concordance between different studies. Using different approaches such as association, family linkage, genome-wide association and meta-analysis, researchers reported numerous genes that might play a role in bone physiology, most of the time searching for correlation with phenotypes such as low bone mineral density (BMD) and fractures. Unfortunately, most of these genetic variations were not further investigated for their functional role and how these could lead to the disease. Some monogenic bone diseases led to the identification of genes that were never considered to be involved in bone physiology such as the low density lipoprotein receptor-related protein (LRP)-5 (Gong et al., 2001) and sclerostin (SOST) genes (Brunkow et al., 2001).

A genome-wide linkage scan was performed in two Maltese families with a very high incidence of osteoporosis, where suggestive linkage to chromosome 11p12 was observed. After investigating the genes known to be found at this region by DNA sequencing, we identified a variant in the CD44 gene that was co-segregating with the inherited haplotype in all affected members within one of the families. Further studies on this variant suggested that it could affect pre-messenger RNA splicing, or organisation, leading to different levels of slightly modified variants of the same protein (isoforms). Other loci were identified in both families.

Without doubt, the analysis of data would not have been possible without the number of bioinformatic tools and software that are available. The advances in computer technology including the internet, led to the development of various software and online tools. In this

chapter, we will take a look at software and other online tools used in this study. We will discuss the basic concepts of the study, how the analysis was performed using different software and the interpretation of results.

1.1 Gene mapping using families

One of the greatest challenges for geneticists is the identification of genes responsible for complex disease. Unlike classical Mendelian disorders, these diseases do not show obvious Mendelian patterns of inheritance and involve complex interactions between various environmental and genetic factors. Confounding factors such as heterogeneity, phenocopies, genetic imprinting and penetrance further complicate the identification of susceptibility genes.
When performing a genetic study, correlation between phenotype and genotype is sought. In complex traits, this correlation might be very low due to incomplete penetrance where not all individuals having the same susceptibility allele are affected or where affected individuals do not have a susceptibility allele (phenocopies). These factors lead a wide ranging severity of disease even within a single family. Further more, late onset diseases such as cardiovascular disease and osteoporosis show up later in life and thus unaffected individuals tested today might become affected in the near future. Late onset diseases are more sensitive to environmental (mostly lifestyle related) factors and are observed to have a higher level of genetic variation due to weak selective pressures on these variants that are usually neutral early in life (Wright et al., 2003). Besides testing for a qualitative trait where individuals are grouped as either having or not having the disease, one can use quantitative or a continuous measurement such as BMD. When using a quantitative variable one must be very cautious as it might not completely correlate with the disease and it could also be dependent on a number of other non-genetic factors including limitations of methodology.
Complex disorders are most often polygenic where multiple genes contribute to the phenotype. Complex patterns of inheritance might be due to allelic or locus heterogeneity where different variants within the same gene are responsible for the disease or where a number of different genes are involved in the same biological process. When studying complex disorders, therefore, one is looking for susceptibility alleles at multiple loci that together increase the individual's risk for the disease. In polygenic traits, penetrance is determined by the genotypes of other loci and therefore it is likely to be low and will vary between individuals. To increase the chance of successful gene mapping, it is important to identify families from probands with extreme phenotypes, earlier age at onset or else to study families from an isolated population with a very high incidence of disease. Wright and colleagues (2003) suggested that it is important to identify genes with the largest contribution to the extremes of the trait and avoid quantitative trait loci (QTL) that have minimal effects on the individual or disease mechanism. Using single extended families from populations that are homogeneous and consanguineous has proven to be a successful approach in localising the genes and novel mutations in type 2 diabetes (Kambouris, 2005). Using one extended family, Kambouris reported similar results to those obtained from previous genome-wide scans using hundreds of individuals (Hanson et al., 1998). This shows that costs and time to identify novel genes responsible for complex disorders can be significantly reduced, when using extended and consanguineous families coming from homogeneous populations.

1.1.2 Linkage analysis

In linkage analysis, the non-independent co-segregation of marker and disease locus is tested in families with multiple affected individuals. Linked alleles (marker with disease-

causing allele) on the same chromosome segregate together more often than expected by chance; i.e. against Mendel's law of independent assortment. Gene mapping of a trait identifies chromosomal loci that are shared among affected individuals and that differ between affected and non-affected family members. Positive linkage can only be obtained for marker alleles inherited together with disease allele on the same chromosome. This is a major limitation for linkage analysis when different disease alleles present at the same locus are on different chromosomes, hence *in trans,* as in a case of coeliac disease (Vidal et al., 2009a). In this study, no evidence of linkage was observed to the human major histocompatibility complex (MHC) locus on chromosome 6, in a family with high incidence of celiac disease. Further investigations showed that this was because inherited risk alleles coding for HLA group *DQ2.2* occur *in trans* and so cannot be detected by linkage.
For a linkage study family members from pedigrees with normal and osteoporotic individuals are genotyped for a set of polymorphic markers either across the whole genome or at specific chromosomal loci, where known candidate genes are located. Genetic linkage is measured by the recombination fraction that is the probability that a parent will produce a recombinant offspring and is dependent upon the distance between loci. The more distant two markers are from each other the higher is the chance that a recombination event occurs between them during meiosis. The recombination fraction theta (θ) ranges from 0 for completely linked markers to 0.5 for unlinked loci. Genetic linkage is measured in centiMorgans (cM), where 1cM represents 1% recombination or θ = 0.01 that is equivalent to 1 million base pairs. So using the recombination fraction one can calculate the physical distance on the chromosome, although recombination rates might vary depending on location on chromosome. Recombination rate is usually lower closer to the centromere. Also these measurements might not be so accurate for longer chromosomal distances where multiple crossovers might occur during a single meiotic event, a phenomenon known as interference. Two mapping functions to convert recombination fraction into map distance are Haldane's, that does not assume interference, and Kosambi's, which assumes interference as 1 - 2θ.

1.1.3 Parametric linkage analysis

Parametric linkage analysis is a statistical approach using the logarithm of the odds ratio (LOD score) to assess the strength of linkage. This is also known as a model based linkage where the mode of inheritance, frequencies of disease and marker loci together with penetrance must be known. The statistic assumes the likelihood (or probability) that a disease and marker loci in a family are not inherited together (θ = 0.5) compared with the likelihood that they are linked over a selected range of recombination fractions (θ range of 0 to 0.5). The LOD score is the base ten logarithm of the likelihood ratio that is calculated for each value of θ. A two point LOD (z) score (between disease locus and marker) is calculated using the following equation:

$$z(x) = \log_{10} [L(\theta=x) \div L(\theta=0.5)] \quad (1)$$

where x is a value of recombination fraction and L is the likelihood.
Significant evidence of linkage is taken at a LOD score of 3.0 or higher and linkage is completely excluded with a LOD score of -2.5. A LOD score of 3.0 corresponds to odds of 1000:1 that means that it is 1000 times more likely that the alternate hypothesis in favour of linkage holds while a LOD score of 3.5 is equivalent to odds of 3162:1. The LOD score can be converted to a chi-square statistic by simply multiplying by 4.6 and calculating a p-value at

1 degree of freedom (df) for ordinary LOD and at 2 df for heterogeneity LOD scores (HLOD), under the null hypothesis (Ott, 1991). The p-values obtained are always divided by 2 for one-sided tests except when calculating p-values for multi-point LOD (MLOD). Using these calculations a LOD score of 3.0 is equivalent to a p-value of 0.0001 while that of 3.6 is equivalent to 0.00002. However, a chi-square derived p-value applies more for large sample sizes and can be underestimated when sample size is too small. Lander and Kruglyak (1995) suggested that linkage must be reconfirmed by other independent investigators where a nominal p value of 0.01 would be required, while they advised caution when reporting LOD scores that are less than 3.0 and so are only suggestive of linkage. In case of suggestive linkage, additional family data would be required before conclusions can be drawn (Lander & Kruglyak, 1995).

LOD scores can be influenced by a number of factors including the phase or whether parental genotypes are known, misspecification of disease and marker allele frequencies, penetrance, heterogeneity and mostly by phenotypic misclassification. Also for more accurate linkage information and to better localize the disease gene, multi-point linkage analysis is preferred over two-point analysis. Statistical analyses in complex pedigrees are carried out using software such as MLINK and GENEHUNTER where the LOD score can also be adjusted for locus heterogeneity (HLOD) (Kruglyak et al., 1996).

Another kind of analysis which is thought to be useful when analysing linkage data for complex traits is the MOD-score. In complex traits both the genetic model and disease allele frequency are very difficult to specify correctly. An incorrect assumption of the genetic model can significantly affect the analysis and can lead to a false negative result. The MOD score is calculated by maximising the LOD score over a number of replicates using different penetrances and disease allele frequencies, to obtain a maximum LOD score using the best genetic model (Strauch et al., 2003). To control type I errors, it was found that a MOD-score of 3.0 should be adjusted by a value ranging from 0.3 – 1.0 where it was proposed that a MOD-score of 2.5 is indicative of suggestive linkage (Berger et al., 2005). MOD-score analysis can be used to determine the best genetic model for those regions indicated by an initial genome scan using ordinary LODs and it can also be calculated assuming paternal or maternal imprinting. When assuming imprinting a heterozygote paternal penetrance is also used with the other three penetrances with a total of four penetrance values. If a low heterozygote frequency is calculated for paternal imprinting, it indicates that maternal genes are preferentially expressed at that locus (Strauch et al., 2005; Berger et al., 2005).

1.1.4 Non-parametric linkage analysis

Since the mode of inheritance for complex disorders is uncertain, evidence of linkage might be missed by using the LOD score method described above. A more appropriate approach is that described by Kruglyak et al (1996) known as a non-parametric linkage (NPL) or a model free analysis. The NPL statistic measures allele sharing among affected relative pairs (ARP) and/or affected sib-pairs (ASP) within a pedigree. By chance it is expected that siblings share zero, one or two marker alleles identical by descent (IBD) with a probability of 0.25, 0.50 and 0.25, respectively. If disease and marker alleles are linked then affected siblings will share these alleles more frequently than expected by chance regardless of the mode of inheritance. Comparison between expected and observed allele sharing between ASPs is then analysed using the chi-square statistic. Highly heterozygous markers, multipoint linkage and genotyping of non-affected siblings when parents are not available help to

increase the sharing information. One great advantage of the NPL statistic is that data from markers on a chromosome can also be evaluated in a multipoint approach using software such as GENEHUNTER which uses the Lander-Green algorithm to calculate IBD distribution (Kruglyak, 1996).

1.2 Phenotype definition, selection of family and population

1.2.1 Phenotype

Phenotype definition is one of the most important factors and should be determined by proper diagnosis or exclusion of other medical conditions that could lead to the same disease. To exclude disease and other factors leading to secondary osteoporosis, individuals were asked to answer a questionnaire and a series of other medical tests were performed. Measurement of bone mineral density (BMD) together with t-scores (number of standard deviations from the mean BMD of a control group of young women at peak bone mass) is the gold standard to diagnose osteoporosis, as recommended by World Health Organisation (WHO). However, this methodology does not show the whole picture partly because bone strength, thus fracture risk, is not completely assessed by measuring bone density. Also, individuals with normal BMD, who might become osteoporotic in ten or twenty years time, could still carry the responsible allele. As discussed above, miss-classification of affected status might seriously affect the results obtained by statistical analysis. To overcome this issue, and unlike other linkage studies for osteoporosis, we used different thresholds of t-scores and z-scores at the lumbar and femoral sites obtained after measuring BMD, to define discreet phenotypes as simply affected or not-affected. Statistical analyses were performed in five different scenarios defining discreet phenotypes using the guidelines suggested by the International Society of Clinical Densitometry (Khan et al., 2004).

1.2.2 Families

Extended families with a number of affected individuals are ideal for identifying variants with higher penetrance but are less frequently found in populations. Development of novel treatments can be targeted to these pathways. Factors such as mode of inheritance, penetrance and disease or allele frequencies together with technical factors such as accuracy of genotyping, all affect power to detect a significant linkage.

1.2.3 Population

The genetic component within a population is strongly affected by its history and demography. The genetic pool of a population is determined by mutations, population admixture as well as by random genetic drift that occurs most often due to catastrophic events that result in a major decrease in population (Wright et al., 1999). Genetically isolated populations (by geography and/or culture), that recently expanded from a very small number of founders with occasional interbreeding with other ethnic groups, are more likely to share haplotypes identical by descent (IBD) over longer genetic distances (Wright et al., 1999).

The present Maltese population, although geographically (but not genetically) isolated, is thought to have expanded exponentially from a much smaller population during the last four hundred years with a possibility of a number of founder effects introduced by admixture with other populations coming from Sicily, the eastern Mediterranean and northern Africa. Founder effects were reported in the Maltese population, including a

mutation (R1160X) found in the NPHS1 gene coding for nephrin that causes nephrotic syndrome (Koziell et al., 2002) and the 68G>A mutation within the quinoid dihydropteridine reductase gene that causes a rare form of hyperphenylalaninaemia and phenylketonuria (Farrugia et al., 2007). The introduction of founder effects and major bottlenecks may increase the chance of creating sub-populations with particularly high allele frequencies when compared to the rest of the population (Heiman, 2005). Significant fluctuations in the population were brought about by emigration of the Maltese in fear of further attacks by the Turks, death by famine or plague. On the other hand, the existence of a relatively frequent disease in an island population does not necessarily always indicate a possible founder effect since this might result from multiple mutations in a single gene or in different genes that could lead to the same phenotype (Zlotogora, 2007).

Genetically isolated populations proved to be very useful for the identification of genes not only in the case of the BMP-2 gene in Iceland but also for a number of other diseases (Styrkarsdottir et al., 2003). More than 15 mutated genes were successfully identified by positional cloning in families from the isolated population of Finland. The Finnish population demographic history was characterised by rapid expansion from a much smaller population with a number of founder effects (Peltonen, 2000). Another island population that proved successful for the identification of a mutation responsible for uric acid nephrolithiasis by linkage was the Sardinian population (Gianfrancesco et al., 2003). Linkage studies in Maltese families resulted in successful identification of rare genetic variants responsible for other human disorders such as coeliac disease (Vidal et al., 2009a), epilepsy (Cassar, 2008) and recently in the identification and confirmation of the role played by the erythroid transcriptional factor KLF1 in hereditary persistence of foetal haemoglobin (Borg et al., 2010).

2. Materials and methods

2.1 Patient recruitment

Two extended families consisting of a total of 27 family members with several individuals having low BMD were recruited for this study. Families were selected through index patients (or probands) referred to the Bone Density Unit, Department of Obstetrics and Gynaecology, St. Luke's Hospital, Malta for an osteoporosis risk evaluation. The proband in Family 1 was a 61-year-old female diagnosed with osteoporosis six years earlier and was known to have a family history of osteoporosis. Five out of seven of her siblings were recruited while the other two were not willing to participate in the study. Osteoporosis was confirmed in all six recruited siblings. All female siblings were osteoporotic at the lumbar spine and one male was osteoporotic at the femoral neck. One sibling had an asymptomatic compressed vertebral fracture. Three daughters of the proband were recruited (age range 33 – 38 years) and all of them were found to have very low BMD for their relatively young age. Their 37-year-old cousin was also found to have very low BMD at both the lumbar (t-score -2.25) and femoral neck (t-score -1.07), and had very low body mass index (BMI) (16.2 kg/m^2). It was not possible to collect blood for DNA analysis from this participant.

The proband in Family 2 was a 55-year-old woman with osteoporosis at the lumbar spine, diagnosed five years earlier. A closer investigation of this family revealed four osteoporotic siblings out of five. Their children were healthy young adults, some of whom had very low BMD relative to their age. The presence of males with low BMD and history of fractures in a severely osteoporotic sibling were good indicators that a genetic factor might be involved.

As already discussed, five different scenarios were tested using thresholds for t-scores and z-scores as previously described (Khan et al., 2004). Osteoporosis for post-menopausal women and men over fifty years of age was defined using a lumbar and/or femoral t-score of less than -2.50 (WHO criteria). Definition of affected status for younger individuals was determined using z-scores of less than -1.0 and less than -2.0 for a more severe phenotype, for scenarios III and IV, respectively. For scenario V, analysis was performed using only affected individuals having femoral z-scores of less than -1.0. In all five scenarios, family members having normal BMD measurements were assumed to have an unknown phenotype. This assumption takes into consideration the possibility that any apparently clinically unaffected individual might actually be affected, thus reducing the chance of obtaining false negative results.

2.2 Genotyping

To perform a successful gene mapping study, a number of polymorphic markers have to be typed in affected and non-affected individuals to identify genes that increase the risk of disease. Different types of genotyping markers were used in recent years and new techniques for typing are constantly being developed to increase efficiency, accuracy and throughput while reducing costs.

2.2.1 Microsatellite genotyping

Short tandem repeats (STRs) or microsatellites are widely distributed in the genome and so are useful tools for genome-wide scans. These tandem repeats can be dinucleotide, trinucleotide or tetranucleotide repeats where polymorphisms are generated by gain or loss of repeats usually as a result of both replication slippage and point mutation. Microsatellites have several advantages for typing, the most important of which is that they are highly polymorphic with a very high heterozygosity (>70%), so making them ideal for use in linkage studies. Another advantage is that they can be very easily typed using PCR techniques where fluorescently labelled primers flanking the polymorphic region are designed. The variable number of repeats creates amplicons of different sizes which can be typed using automated sequencers such as those by Applied Biosystems (ABI) (PE Applied Biosystems Division, Foster City, CA). Different sets of markers across the whole genome are electronically available from databases such as those of Marshfield Institute of Genetics (http://research.marshfieldclinic.org/genetics/), deCode (http://www.decode.com/services/microsatellite-genotyping-genome-wide-scans.php) and the Cooperative Human Linkage Centre (http://gai.nci.nih.gov/CHLC/). Markers can be selected from these databases either across the whole genome or at candidate loci usually with an average spacing of 10cM and for a higher resolution at < 5cM. To increase throughput and reduce costs, the amplified fragments are carefully pooled in sets in such a way that the allele size range does not overlap within a set and by using different dyes for different sets.

An initial genome-wide scan, 400 microsatellite markers spread across the 22 autosomes and x-chromosome with an average spacing of 8.63cM and heterozygosity of 0.77, was performed. The average performance of markers for all samples was of 96.96%. Fine-mapping was performed by increasing the markers at indicated loci from the initial scan. Genotyping was performed by polymerase chain reaction (PCR) followed by fragment analysis using a 3730xl ABI genetic analyser (Applied Biosystems, Foster City, CA, USA). The average performance of the markers was of 96.02%. Genotyping was performed

commercially at the McGill University and Genome Quebec Innovation Centre, Quebec, Canada.

2.3 Analysis of linkage data

PedCheck (O'Connell and Weeks, 1998) was used to determine if the inheritance of marker loci was according to Mendel's laws. Multipoint parametric and non-parametric linkage analyses were performed using GENEHUNTER-PLUS (Markianos et al., 2001) which is an improved version of GENEHUNTER (Kruglyak et al., 1996). GENEHUNTER v1.2 was used to calculate Zlr scores according to Kong and Cox (1997). Linkage analysis of markers on the X-chromosome was performed using a specific application for this chromosome included with the GENEHUNTER package. All analyses were performed using EasyLinkage v5.05 (http://www.uni-wuerzburg.de/nephrologie/molecular_genetics/molecular_genetics.htm) (Lindner and Hoffmann, 2005). Parametric analysis was carried out using variable penetrances for both a dominant and recessive mode of inheritance. Penetrances used for the dominant model were 0.01 for the wild-type homozygote, 0.90 for mutant heterozygote and 0.90 for mutant homozygote, respectively. The recessive model was defined by penetrances 0.01, 0.01, and 0.80 for the wild-type homozygote, mutant heterozygote and mutant homozygote, respectively. A more complex model was also analysed using penetrances 0.01, 0.05, 0.30 for wild-type homozygotes, mutant heterozygotes and mutant homozygotes, respectively. A parametric analysis assuming heterogeneity was computed using data from both families (HLOD).

A co-dominant allele frequency algorithm was used for the analysis, as suggested in the EasyLinkage manual, for extended families. For all models, the disease allele frequency assumed was 0.001, and phenocopy rate of 1%. This disease allele frequency is equivalent to a population prevalence of 0.2% assuming Hardy-Weinberg equilibrium calculated using the following equation (Xu & Meyers, 1998):

$$2(1 - q)\, q + q2 \tag{2}$$

q = disease allele frequency.

Analysis was performed using other penetrance values for loci showing evidence of linkage in the initial genome-wide scan. The exact genetic model was determined using GENEHUNTER-MODSCORE v1.1 (Strauch et al, 2005), where MOD scores were calculated from simulations of different models and disease allele frequencies with and without imprinting. This analysis was suggested by Strauch et al (2003) for complex trait analysis and was done only for those regions showing suggestive linkage. The deCode genetic map was used throughout the study.

2.3.1 Using EasyLinkage v5.05 graphical user interface (GUI)

EasyLinkage is a Microsoft Windows® based GUI, developed in recent years. This was a step forward for researchers wanting to perform linkage analysis. Using EasyLinkage and a common input file format, one can analyse data using all major software such as PedCheck, GENEHUNTER, Merlin and Allegro. EasyLinkage can be used to analyse data generated from projects using both single nucleotide polymorphisms (SNPs) as well as STRs. Analysis can be performed on chromosome by chromosome or else genome-wide basis, making use of the appropriate genetic maps (such as deCode and Marshfield), using male, female or sex-averaged maps, from which more accurate genetic positions can be drawn. Both

graphical and text output files are automatically generated for each individual family together with a collective report averaging all families, in text or pdf formats and stored into an appropriately labelled folder showing date and type of analysis. These files show statistical analyses results such as LOD scores, NPL, p-values and input parameters given by the user for that model including penetrances, disease allele frequencies and genetic positions of markers ranked according to the most significant results.
There are four allele frequency algorithms to choose from depending on the type of analysis needed. Several versions of this software have been developed, improving its capabilities to handle large amounts of data generated from SNP arrays such as the Affymetrix 500k and Illumina 650k chips. For SNP analysis, allele frequencies of all the major ethnic groups form part of the EasyLinkage software package.

2.3.2 Data entry

There are two main types of files needed to perform linkage analysis using microsatellites or STRs. In this study, a qualitative type of analysis was performed using discrete phenotypes (affected vs unaffected). One type of input file should contain family or families' information in a standard linkage format. The marker file should include the genotype results for each family member. All family or families' information including pedigree structure has to be entered into a pedigree file. Shown below is part of the pedigree file as created in our study (only obligatory columns were used). From left to right columns represents (i) unique family identifier; (ii) individual unique identifier (iii) father and (iv) mother identifiers; (v) sex identification code (1=male, 2=female, 0=unknown); (vi) affected status (1=unaffected, 2=affected, 0=unknown). In case parents are unknown then enter '0`. As explained previously is an unknown phase and so it reduces the power of the study, even though the software is able to assume the genotypes of these individuals using the known genotypes from their offspring (inferred genotypes). An example for using the unknown option in column (v) is when you do not know the sex of a child due to death *in utero*.

A_1	A_1_01	A_1_11	A_1_12	2	2
A_1	A_1_02	A_1_11	A_1_12	1	2
A_1	A_1_03	A_1_11	A_1_12	2	2
A_1	A_1_10	0	0	1	0
A_1	A_1_11	0	0	1	0
A_2	A_2_11	A_2_28	A_2_29	1	1
A_2	A_2_12	A_2_28	A_2_29	2	2

Phenotype definition has to be done using appropriate criteria and diagnostic tests, for example in our study, measurements of BMD together with blood tests were used to exclude other medical conditions that could also affect BMD. In complex disorders, it might be difficult to define the phenotype correctly and this could seriously affect the outcome of results. Select 1 and 2 wherever diagnostic tests were performed and phenotype is known. Any individuals not tested should be defined as having an unknown status. This is a better option because individuals wrongly defined as normal could give a false negative result (type II error) as these might be carrying the causative alleles and might become affected at a later stage in their life. As described in the phenotype definition section, we analysed our data using five different scenarios defined by t-scores and z-scores. For each scenario a

different pedigree file was created and saved in a folder together with the marker files described in the next section.

Creating marker files

Marker files should include genotyping results for all family members tested. Entering data is the most laborious part of the study because different files have to be created for each marker, i.e. if 400 markers were tested then 400 different marker files have to be created and saved in the same folder together with pedigree files. These files should be named with the marker identification corresponding with that in the marker map file (.map) used by the software e.g. 'D1S200_FINAL.abi`. If the marker is not found within the marker map file then an error is given when running the analysis. This error can be corrected manually by adding the marker into the marker map file found in the EasyLinkage folder in Program Files.
This is an example of the method used to input data into marker files:

MARKER	LANE	ID	A_1	A_2
D1S200	A1	A_1_01	165	176
D1S200	A2	A_1_02	161	176
D1S200	A3	A_1_03	165	176
D1S200	A4	A_1_04	161	176
D1S200	A5	A_1_05	161	176
D1S200	A6	A_1_06	161	176
D1S200	B1	A_2_10	161	176
D1S200	B2	A_2_11	161	176
D1S200	B3	A_2_12	161	161

Column (i) name of marker e.g. D1S200; (ii) PCR reaction position in a 96-well PCR plate (information not used by software); (iii) individual identification number corresponding with pedigree file; (iv) allele 1 in base pairs (bp) and (v) allele 2 in bp. Any missing genotypes should be entered as '0`. When analysing data, the software will re-code these alleles numbering them consecutively as 1, 2 etc depending on the number of alleles observed for that marker in all genotyped individuals. The higher the number of alleles observed the higher the heterozygosity and thus the more informative that marker is.

2.3.3 Running EasyLinkage analysis

On the main screen of the GUI, we selected a 'Single Locus` analysis, the linkage software (GENEHUNTER) and microsatellites project type. Next step was to select whether to perform a genome-wide analysis, one chromosome at a time or even to analyse small segments from a chromosome. Analysing small segments from a chromosome is useful to analyse large scale SNP data possibly analysing 500 markers in one segment. A lower LOD score was observed when analyzing a large number of markers, which would mean that for SNP analysis, it is better to avoid SNPs that are very close to each other. LOD scores were observed to be lower in such instances most likely due to allele frequencies used. It would be advisable to first analyse the whole chromosome for SNP analysis, and if significant results are observed, then re-analyse blocks of 100 markers at a time and as overlapping blocks. Another suggestion would be to use different and appropriate allele frequency algorithms, as will be described below.

After choosing the chromosomes, the sex-averaged deCode genetic map was selected to position the markers. There is a difference of approximately 10 cM between the male and female genetic maps, being longer in females due to a higher recombination rate. Other general options selected included 'recode alleles` for continuous recoding of alleles within the marker files, Mendelian testing using PedCheck and the autoscale Y-axis for LOD / NPL plots.
Finally we chose the folder where the pedigree files were saved and the option to give individual pedigree results as well as totals. As decribed earlier, five different phenotype scenarios were used and each one had to be analysed using a different pedigree file.

GENEHUNTER (GH)

This computer package was developed to perform multipoint linkage analysis (parametric and non-parametric) in pedigrees of moderate size (Kruglyak et al., 1996). The program can compute LOD scores for pedigrees using a mode of inheritance and penetrance specified by user. It also allows the user to test for linkage under genetic heterogeneity. The multipoint NPL analysis tests for IBD allele sharing among affected individuals within pedigrees that is not affected by the mode of inheritance. It is thus ideal to be used for complex traits. GH also constructs the most likely haplotypes indicating crossovers even if there is missing data. A major advantage of GH over other statistical software, such as VITESSE and MLINK, is that it uses the Lander-Green algorithm and therefore it can perform multipoint analysis using several markers on a chromosome. Major drawbacks of GENEHUNTER include restrictions on pedigree size and its relatively slow speed when compared to similar software such as Allegro. Another limitation of GENEHUNTER is that it cannot analyse large number of markers which means that if one was analysing more than 100 markers on same chromosome, one would have to analyse these in groups of 100, repeating the analysis with different set sizes so as not to miss the signal. A recent version of GH can also perform transmission disequilibrium testing (TDT) analysis and analysis of quantitative traits making GH the ideal software to use for genetic analysis (Nyholt, 2001). In this study GENEHUNTER v1.2 was used to calculate Zlr scores using the Kong and Cox (1997) model. This algorithm addresses the problem encountered by previous versions of GENEHUNTER where NPL scores were found to be too conservative when inheritance data was incomplete. Another application used in this study was GENEHUNTER-MODSCORE v1.1 that maximises LOD scores with a series of penetrances and disease allele frequencies (Strauch et al., 2005).

Using GENEHUNTER (GH) with EasyLinkage GUI

Performing linkage analysis using GH through the Easylinkage GUI is easy and straightforward and saves time. After choosing the GENEHUNTER package software as described above, one has to go to Program Options from the main dashboard to be able to define a model for analysis. We analysed our data using both dominant and recessive models of disease. GENEHUNTEr v1.2 and GENEHUNTER-MODSCORE v1.1 were used for the analysis. A 'Codominant` allele frequency algorithm was used for our analysis. EasyLinkage gives you four different algorithms to choose from. The Codominant algorithm was the best choice to use for extended families. This algorithm uses only alleles from genotyped individuals within the pedigree file. Frequencies of the alleles are calculated to sum up to 1, which means that if 5 different alleles were observed, then the allele frequency for each allele will be set to 0.200 or if 10 then to 0.100. If less than 5 alleles are found then still the frequency is set to 0.200. Other allele frequency algorithms include either all

individuals within the marker file or all individuals from pedigree file, both suitable for the affected sib pair (ASP) design. There are also specialised algorithms such as 'founders only`, suitable only for pedigrees with large number of founders. SNP projects will use reference allele frequencies from different ethnic groups.
As described before, we analysed our data using variable penetrances of disease starting with a highly pentetrant form (90%) down to 50%, for each scenario. Disease allele frequency was taken as 0.001 and the analysis steps between markers for multi-point analysis were set to 5, with recombination counting set to 'On`. Penetrances were entered into the appropriate fields as described before, turning the haplotyping options to 'on` and choosing the 'Display all family plots`. The haplotyping option significantly increases the run time of the analysis but it creates plots for each family with haplotypes and marker positions together with other files that can be used by other software such as HaploPainter (Thiele & Nurnberg, 2004).

3. Results

3.1 Reading analysis files

All analysis files were saved into an appropriately labelled folder with details of software used, allele frequency algorithm, pedigree file name, date and time of analysis. Data was saved as text and post-script format. The '.OUT` file within the 'LOG` folder can be opened using Notepad where one can find all commands given to GENEHUNTER by EasyLinkage. If any errors were encountered and analysis was not completed, then one would find all information logged in, within this file. This file also includes detailed results such as LOD / NPL scores, marker information etc for each individual family and totals for all the families. EasyLinkage automatically commands GENEHUNTER to set up the maximum number of bits to be analysed within the family and to 'trim` large families when needed. These two functions are needed to keep computations within the computer running ability due to memory limitations. Max Bits function is dependent on the number of meiosis being examined and represents 2N – F, where N is the number of non-founders and F is the number of founders. If, in the family there are 10 children (non-founders) from 2 sets of parents (4 founders), then max bits will be set to 16. If the family is larger, then the less informative individuals will be excluded from the analysis using the 'trim` command. Automatically EasyLinkage also commands GENEHUNTER to use the Haldane map and there is no option to use the Kosambi map. The Haldane map does not assume interference, as described in a previous section.
Figure 1A, shows the upper left hand quandrant of a parametric LOD score plot, including information such as allele frequency algorithm used, inheritance model, allele frequencies and penetrancies expressed as percentages. Markers are ranked in order of their most significant LOD / NPL scores with information about chromosome number, their position on the genetic map and calculated probability (Figure 1B). Figure 1C shows NPL results for a genome-wide analysis with NPL scores on the y-axis and chromosomal marker positions on the x-axis. Other plots are given for LOD scores, HLODs, p-values, Zlr scores and marker information.
If the haplotyping option was chosen then a folder named 'Haplotyping` is created and haplotype files for each pedigree are saved there. This haplotype data can then be imported to other software such as HaploPainter (Thiele & Nurnberg, 2004) to construct a graphical representation of the inherited haplotypes.

A

Project:	OSTEOPOROSIS	Inheritance:	Dominant
Family name:	A_1	Common allele:	99.90 %
Used map:	decode strp[1] (sex averaged)	Disease allele:	0.10 %
Marker positions:	24 ok / 0 ? / 0 outside	LC1 PCOPY rate:	1.00 %
Allele frequencies:	Equal distribution of alleles	LC1 PENET wt/mt:	90.00 %
HAP algo / SCFct:	On / All	LC1 PENET mt/mt:	90.00 %

B

Marker	CHR	cM	pLOD	NPL	npLOD	P val	INFO	Alpha	hLOD
1. D6S309	6	20.50	0.4256	−0.3697	0.0297	0.7188	0.9222	–	–
2. D6S1006	6	30.18	0.4235	−0.3736	0.0303	0.7188	0.9201	–	–
3. D6S309 – D6S1006	6	22.44	0.4230	−0.3694	0.0296	0.7188	0.9026	–	–
4. D6S1713 – D6S309	6	18.66	0.4229	−0.2391	0.0124	0.4375	0.6361	–	–
5. D6S309 – D6S1006	6	28.24	0.4217	−0.3717	0.0300	0.7188	0.9016	–	–

C

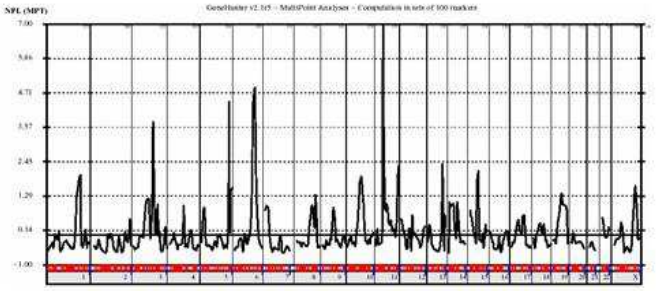


Fig. 1. Information given by parametric LOD score analysis plots (A & B) and genome wide NPL plot (C)

3.2 Linkage results

From the initial genome-wide scan using both families, evidence of linkage was observed to marker D11S1392, where the highest NPL score was of 5.77 (p=0.0006) and LOD/HLOD of 2.55, for the dominant model with 90% penetrance and phenocopy rate of 1%. Fine mapping was performed by analysing four additional markers at this region (D11S4101, D11S935, D11S4102, and D11S1911) with average spacing of 1 to 1.5cM. Fine mapping confirmed linkage to marker D11S935 that is 52.94cM from 11p-telomere. Table 1 shows the highest scores obtained for this marker using the dominant mode of inheritance with 90% penetrance and phenocopy 1%. HLODs are calculated when more than one family are analysed together and thus the score can be different from the LOD if there is heterogeneity between families.

Table 2 shows results obtained when analysing the same families assuming clinically unaffected individuals, whose BMD was measured by DEXA, as normal phenotype (according to WHO criteria) rather than having unknown phenotype (as in Table 1). NPL and Zlr scores were the same as observed in Table 1, but LOD and HLOD scores were different.

LOD scores shown are for the autosomal dominant model with 90% penetrance and 1% phenocopy rate.

Phenotype	LOD (cM)	HLOD (α)	NPL (p-val)	Zlr
Scenario I	2.90 (52.94)	2.90 (1.00)	7.00 (0.0014)	3.01
Scenario II	2.46 (52.94)	2.46 (1.00)	4.02 (0.0038)	2.90
Scenario III	2.89 (52.94)	2.89 (1.00)	7.23 (0.0013)	3.04
Scenario IV	3.35 (52.94)	3.35 (1.00)	6.90 (0.0002)	3.74
Scenario V	2.59 (52.94)	2.59 (1.00)	5.37 (0.0020)	3.28

deCode map position in brackets in cM

Table 1. Highest scores for marker D11S935 in both Pedigrees using an Autosomal Dominant Model after Fine Mapping

This is because in the second analysis, shown in Table 2, based on current BMD measurements were defined as normal, individuals that might be osteoporotic in the future. These individuals might also be carrying the inherited causative allele and so will result in a false negative result if taken as normal. This is a common situation with complex and late onset disorders such as osteoporosis.

Phenotype	LOD (cM)	HLOD (α)	NPL (p-val)	Zlr
Scenario I	3.07 (52.94)	3.07 (1.00)	7.00 (0.0014)	3.01
Scenario II	-0.19 (52.94)	-0.00 (0.00)	4.02 (0.0038)	2.90
Scenario III	2.97 (52.94)	2.97 (1.00)	7.23 (0.0013)	3.04
Scenario IV	2.80 (52.94)	2.80 (1.00)	6.90 (0.0002)	3.74
Scenario V	1.26 (50.64)	1.26 (1.00)	5.24 (0.0020)	3.28

Table 2. Analysis of Chromosome 11 in Both Families Assuming Unaffected Individuals as Normal

When calculating MOD scores for chromosome 11, the highest MOD was of 3.28 at the same region 52.94cM using the best calculated genetic model with penetrances 0.06 wild-type homozygotes (6% phenocopy rate), 0.97 for both heterozygotes and mutant homozygotes. The disease allele frequency calculated at this model was of 0.000006 with a calculated population prevalence of 0.001%, assuming Hardy-Weinberg equilibrium. A MOD score of 4.33 (info = 0.87) was obtained when assuming imprinting with a disease allele frequency of 1×10^{-6}. Estimated penetrances of wild-type homozygote (f +/+) 0.00; paternal heterozygote (f m/+) 0.00; and 1.00 for both maternal heterozygote (f +/m) and mutant homozygotes (f m/m) show evidence of paternal imprinting at this locus. Paternal imprinting indicates that the expression of the gene responsible for the disease at this locus may be entirely maternal.

This locus was further analysed by varying the penetrance and phenocopy rates for the dominant mode of inheritance. Analyses were performed using phenocopies from 1% to 20% and penetrance 0.7 – 0.5. The phenocopy rate is the percentage of individuals within the family that are clinically affected but do not carry the disease allele and hence their phenotype is due to other mainly environmental factors. As shown in Table 3, the highest LOD/HLOD score (3.32) was observed at penetrance of 0.8 and 0.7 with a 5% phenocopy. Changing the penetrances and hence the model, does not affect NPL scores (since these are model free) and therefore NPL scores are not shown in Table 3.

	Family 1			*Family 2*		
	LOD (cM)*	***NPL (p-val)***	***Zlr***	***LOD (cM)****	***NPL (p-val)***	***Zlr***
Scenario I	1.92 (54.35)	6.26 (0.0078)	2.84	1.04 (52.48)	4.42 (0.0098)	2.41
Scenario II	1.35 (55.77)	3.10 (0.0313)	2.12	1.18 (52.94)	3.06 (0.0625)	2.11
Scenario III	1.92 (54.35)	6.26 (0.0078)	2.84	1.04 (51.56)	4.74 (0.0117)	2.27
Scenario IV	1.64 (54.35)	4.41 (0.0156)	2.58	1.77 (52.94)	5.85 (0.0156)	2.75
Scenario V	0.86 (48.21)	1.94 (0.1250)	1.63	1.75 (50.64)	5.83 (0.0156)	2.88

Table 3. Multipoint LOD/HLOD Scores on Chromosome 11 under an Autosomal Dominant Model with Variable Penetrance and Phenocopy

Although both families shared the same linkage interval, the highest LOD scores were obtained by two different markers with a spacing of approximately 4cM between them, showing also different inherited alleles, suggesting that different genes at the same locus, and within the same linkage interval, might be responsible for the same disease in different families (allelic heterogeneity). Highest LOD and NPL scores (1.77 and 5.9, respectively) were obtained for marker D11S1392 (50.64cM) for Family 2, while for Family 1 highest scores were obtained to marker D11S4102 (54cM) (Table 4). Inherited haplotypes identical by descent were observed in both individual families between markers D11S1392 and D11S935, with a number of recombination events defining boundaries for the linkage interval where the causative genes can be found in between.

Penetrance	**LOD/HLOD Phenocopy = 1%**	**LOD/HLOD Phenocopy = 5%**	**LOD/HLOD Phenocopy = 10%**	**LOD/HLOD Phenocopy = 15%**	**LOD/HLOD Phenocopy = 20%**
0.9	3.07	3.25	3.05	2.78	2.47
0.8	3.10	3.32	3.11	2.82	2.44
0.7	3.12	3.32	3.08	2.73	2.23
0.6	3.14	3.30	2.99	2.54	1.80
0.5	3.17	3.25	2.84	2.17	1.07

Table 4. Analysis of chromosome 11 for Families 1 & 2 after fine mapping * deCode map position in brackets in cM

3.3 Choosing and sequencing candidate genes

The locus indicated by both parametric and non-parametric linkage analyses on chromosomes 11p12 was scanned for known candidate genes. Candidate genes within the linkage interval were selected with prior knowledge of physiology using the NCBI map viewer (http://www.ncbi.nlm.nih.gov/mapview/) *Homo sapiens* build 36. The online application GeneSeeker v2.0 (http://www.cmbi.kun.nl/GeneSeeker) was also used. A new online tool GeneDistiller (http://www.genedistiller.org/) was recently developed to filter genes within a specified linkage interval is (Seelow et al., 2008). When using NCBI

MapViewer to select the genes manually, one has to align the genes with the corresponding genetic map (e.g. deCode), setting the resolution of the map to 1 cM for accurate alignment. Applications such as GeneDistiller can facilitate this process since they automatically extract all genes within a given interval. It is also advisable to search for genes further away from both ends of the linkage interval even up to 5 – 10 cM. This will compensate for differences in positioning of markers on the genetic map and the actual physical map.
The whole area from 49 to 55cM on chromosome 11 (deCode genetic map) was searched for genes that might plausibly be involved in the disease. More than twenty genes and hypothetical proteins are found in this region, with the best candidates being the tumour necrosis factor receptor-associated factor 6 (TRAF6) gene [MIM 602355] and the CD44 gene [MIM 107269] (sequenced in Family 2) found 1cM away from D11S1392 (~51cM). TRAF6 was sequenced in both families but it was closer to D11S4102 showing highest scores in Family 1.
Oligonucleotide primers were designed using the online application Primer 3 (http://frodo.wi.mit.edu/primer3/) (Rozen and Skaletsky, 2000) to amplify all coding regions including intron-exon boundaries and promoter region of the CD44 gene using transcript ENST00000278385) from the ENSEMBL database (http://www.ensembl.org). Transcript ENST00000313105 was used for TRAF6 gene. Due to limitations of the sequencing technique, only fragments from 200 to 600bp were amplified by PCR. Large exons and up to 1500bp of the 5` untranslated region were covered by overlapping PCR fragments. Bidirectional DNA sequencing was performed using standard techniques and fluorescent capillary electrophoresis.

3.4 Reading and Interpretation of sequencing results

DNA sequencing results were compared to reference sequences in public databases by using the software ChromasPro v1.33 (http://www.technelysium.com.au) that directly searches the BLAST application on NCBI. Variations that did not agree with the reference sequence were confirmed by the reverse sequence. Electropherograms were also printed and checked manually.
Detailed information about specific genes including information about known mutations/polymorphisms and gene expression was obtained from GENECARDS (http://www.genecards.org) and The Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/). When identifying a variation, the first step is to check using databases whether it is already known. Information about individual SNPs can be searched in gene and SNP databases such as the NCBI SNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/). A list of SNP databases can be found at http://www.humgen.nl/SNP_databases.html. If the variation is a known SNP, then one has to refer to it using the database reference number such as ‘rs3830511`. In the database one can find information about individual SNPs including any population frequencies. The next step is to identify the frequency of these variations in the local population and determine whether it is a rare or common variation. To perform population screening, one can use techniques such as restriction fragment length polymorphism (RFLP), real-time PCR or direct DNA sequencing in a random sample from the general population. For our studies, random samples of DNA were obtained from newborns and used for this purpose, followed by a small scale case-control study using osteoporotic and normal post-menopausal women.

Virtual restriction fragment length polymorphism (RFLP) gel electrophoresis was carried out using the online web applications NEBcutter V2.0 (http:// tools.neb.com /NEBcutter2/index.php) to test the identified polymorphisms.

3.4.1 TRAF6 Sequencing and functional assays

Following direct sequencing, three different variants were identified when compared to reference sequences on the NCBI and Ensembl databases. An A to T transversion was identified at position -721 (5` upstream of exon 1), when compared to TRAF6 reference sequence (AY228337). This variant had not been previously described. Following sequencing of all family members, three affected individuals from Family 1 were observed to be heterozygous for this variant. Individuals from Family 2 were all wild-type homozygotes. RFLP was carried out in 82 unrelated postmenopausal women. This variant was observed to be very rare within an unrelated group of postmenopausal women, as only three heterozygotes were observed. After screening 350 chromosomes in a random sample from the general population, only 2 alleles were observed (0.57%) with this variant having a population frequency of 1.1%.

A previously described insertion/deletion of a T in the intron between exons 4 and 5, in the polyT region, sixteen base pairs ahead of the exon-intron boundary (rs3830511), was also identified. When analysing all family members and controls, only three individuals were observed to be heterozygotes for this insertion/deletion, one of whom was severely affected and the other two were normal individuals.

A transition from G to A was found in the intron between exons 6 and 7, 110bp upstream of the exon-intron boundary. When sequencing all members from both families, four heterozygotes for this variant were identified, and the rest were homozygous for the wild type allele G. Three of the four heterozygous individuals for this variant had a low BMD, two of whom were also heterozygous for the T insertion/deletion described above. Genotyping by RFLP (PvuII) was performed in 82 unrelated postmenopausal women. Genotype frequencies observed were 72.3% GG, 26.5% GA and 1.2% AA.

Although the -721 A/T polymorphism was not linked to the inherited haplotype within Family 1, this polymorphism was rare within the population and it was thus hypothesised that it could affect gene expression. The TRAF6 gene plays a major role in osteoclast differentiation and activation and plays an important role in osteoimmunology (ref).To test this hypothesis the TRAF6 gene promoter region, harbouring the -721 variant, was analysed for possible transcriptional factor binding sites in the presence and absence of the variant identified in this study, using the online application MatInspector by Genomatix Software GmbH(http://www.genomatix.de/online_help/help_matinspector/ matinspector_help.html) (Cartharius et al., 2005). The whole sequence, up to 1500bp upstream from the transcriptional start, site was thus copied and tested using the MatInspector online application. Free registration was needed to use this application for academic use allowing twenty analyses per month. Both normal and mutated sequences were used and analysis was performed using a vertebrate matrix. When comparing normal and mutated alleles it was observed that position -721 might be a potential binding site for nuclear factor Y (NF-Y), a CCAAT binding factor, that binds to the wild-type allele but not to the mutated one. Non-binding of this factor would result in a decreased expression of the gene.

Three different sized fragments from the promoter region of TRAF6 (up to 1500bp) were cloned into a luciferase reporter vector and transfected into two types of mammalian cells. After measuring luciferase activity in both cell lines, it was evident that gene expression was affected by the -721 variant found in the TRAF6 gene promoter. Expression of the mutated allele was observed to be as low as 5% that of the normal allele expressed in murine macrophages. The two longer constructs showed higher expression for the mutated allele suggesting that other transcriptional factors most likely interact either directly or with other factors at the mutated site. Although these observations suggest that this variant affects a transcriptional factor binding site and thus could increase the risk for osteoporosis, further research is needed to identify the molecular mechanisms.

3.4.2 CD44 gene sequencing in family 2

DNA sequencing of the CD44 gene found on chromosome 11p12 was performed in Family 2 since this gene is found closer to D11S1392, which shows the highest LOD scores within this family, as described above. Osteoclast formation was inhibited by CD44 antibody suggesting its important role in bone physiology and as a potential therapeutic target for metabolic bone disease (Kania et al., 1997). As well, CD44 was also associated with inflammatory bone loss (Hayer et al., 2005).

Sequencing CD44 revealed a number of intronic sequence variants, including two A/G changes (rs4756196 and rs3736812) and an A/C transversion in intron 16, none of which were observed to be inherited with the linked marker. A number of other variants were found in coding regions, including an A/G (rs9666607) and C/T (rs11607491) changes in exon 10, both resulting in an amino acid change, which were not linked with the inherited haplotype. Another C/T synonymous variant (rs35356320) was detected in three affected and one unaffected individual. A non-synonymous variant found in exon 12 (rs1467558) was only found in two affected members of this family.

An interesting variant was detected in exon 9, a synonymous G/A transition (rs11033026), found 32 nucleotides upstream from the exon/intron junction. Sequencing the gene in all members of this family, revealed that all individuals carrying the linked STR allele 3 (Figure 1B) for marker D11S1392 were also heterozygous for this variant, suggesting that the two were linked.

As shown in Figure 1B, all affected members, with the exception of one phenocopy (III:4), were heterozygous for both the STR allele and the A allele. This variant was not found in any of the non-affected family members, with the exception of two who also carried the linked STR allele (III:2 and III:7) (incomplete penetrance). The minor allele frequency within the Maltese population was determined to be 0.012 (1.19%) with a population frequency 2.38%. According to the NCBI dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=11033026), and HapMap, this frequency compares with that found in Sub-Saharan Africans, African-Americans and Han Chinese from Beijing (minor allele frequencies 0.336, 0.115, 0.012, respectively), and was absent in European Caucasians. This suggests a founder effect in the Maltese population, complementing other previously reported studies on other human diseases (Farrugia et al., 2007; Koziell et al., 2002).

Since this variant was found in an exon but does not result in an amino acid change, we hypothesised that it could affect pre-mRNA splicing resulting in a different protein isoform.

To test this hypothesis at the transcriptional level an online Bioinformatics predictive tool was used to identify any possible exon splicing enhancers (ESEs) at this region as described by Cartegne et al (2002) (http://rulai.cshl.edu/tools/ESE2/index.html) . The G/A variant was found to abolish an ESE motif (TGAGGA > TGAAGA) for the SR protein (SRp55) with a score of 2.817 (threshold 2.676), in the presence of the A allele. Another online application RESCUE-ESE (http://genes.mit.edu/fas-ess/) did not predict any possible ESEs at this locus (Wang et al., 2004).

The experimental approach to test this hypothesis involved the use of an *in vitro* exon-trapping vector where the whole exon 9 and adjacent introns were inserted into a vector yielding a hybrid construct made up of two vector β-globin exons flanking CD44 exon 9 and adjacent introns. Following transfection into mammalian cells, the construct was transcribed under the control of a SV40 promoter and spliced. The mRNA derived from this construct was extracted and reverse transcribed followed by specific amplification using cDNA as template and specific primers to β-globin exons (SD6 and SA2). The spliced transcripts were analysed by agarose gel (Vidal et al., 2009). Our results showed that in the presence of the A allele only one transcript (261bp) was weakly amplified in both COS-7 and HeLa cells and was completely absent in RAW264.7 macrophages. DNA sequencing confirmed that this transcript did not contain any part of CD44 exon 9, and was entirely made up of vector exon sequences, suggesting skipping of exon 9. Two transcripts were amplified in the presence of the G allele (378bp and 261bp).

4. Conclusion

In this study, two polymorphisms with a population frequency of less than 5.0%were identified by linkage analysis in two extended Maltese families with a highly penetrant form of osteoporosis. *In vitro* functional studies confirmed that these polymorphisms might increase the individual's susceptibility to osteoporosis. This study adds to the existent knowledge of the complex pathophysiology involved in disorders such as osteoporosis. This knowledge is useful for the development of more targeted and individualised treatments. Our results added to the increasing evidence that rare but functional polymorphisms are also responsible for disorders such as osteoporosis, and also that using extended families with extreme phenotypes increases the chance to identify the responsible genes. Computer technology and the internet contribute significantly to the outcome of these studies. Both technologies were important tools for researchers throughout the whole study starting from planning and design of experiments, analysis of data and interpretation of results.

5. Acknowledgements

We would like to thank the families that participated in this study for their collaboration and Dr Raymond Galea for his support in patient recruitment. We also thank Dr Andrew Verner and the staff at the Genotyping Facility McGill University and Genome Quebec Innovation Centre, Montreal Canada for the STR genotyping and Dr Marisa Cassar at MLS BioDNA Ltd, Malta, for DNA sequencing of genes. We would like to thank Dr. Pierre Schembri Wismayer M.D., Ph.D. and Dr. Anthony Fenech Ph.D., B.Pharm. (Hons.), for the use of facilities at the Department of Anatomy and Cell Biology, and Department of Pharmacology and Clinical Therapeutics, University of Malta, and Prof Junko Oshima, Department of Pathology, University of Washington, Washington, USA, who generously

donated the pSPL3 plasmid. This project was approved by the Research Ethics Committee, and supported by the Research Fund Committee of the University of Malta.

Glossary

Allele	Alternative states of genes only identical if their base sequences are identical
Body Mass Index (BMI)	A statistic of the relationship between weight and height = body weight divided by height squared
Bone mineral density (BMD)	A measure of bone density usually measured by x-ray techniques
Haplotype	A set of variants (SNPs or STRs) that are inherited together as a single block on a linear chromosome
Imprinting	Expression of genes depending upon the parent of origin
Linkage disequilibrium (LD)	Groups of markers or genes on the same linear chromosome that are inherited together more often than expected by chance as long as genetic recombination does not take place between them. LD can be used to locate genes associated with phenotype
Locus Heterogeneity	Variability of chromosomal regions involved between different subjects
Penetrance	The percentage of individuals that express a trait determined by gene/s
Phenocopy	A phenotyping change that mimics the expression of a mutation usually resulting from effects of the environment
Segregate	Separation of homologous chromosomes at random during meiosis
SNP	A difference in a single nucleotide at a particular DNA site
STR	Short tandem repeat variations differing between different individuals in the number of repeated sequences eg: (CACACA) or (CACACACACA). Used as markers in forensics for identification.

6. References

Berger, M., Mattheisen, M., Kulle, B., Schmidt, H., Oldenberg, J., Bickeboller, H., Walter, U., Lindner, T.H., Strauch, K., & Schambeck, C.M. (2005). High factor VIII levels in

venous thromboembolism show linkage to imprinted loci on chromosomes 5 and 11, Blood, Vol.105, No.2, (January 2005), pp. 638 – 644, ISSN 0006-4971

Borg, J., Papadopoulos, P., Georgitsi, M., Gutiérrez, L., Grech, G., Fanis, P., Phylactides, M., Verkerk, A.J., van der Spek, P.J., Scerri, C.A., Cassar, W., Galdies, R., van Ijcken, W., Ozgür, Z., Gillemans, N., Hou, J., Bugeja, M., Grosveld, F.G., von Lindern, M., Felice, A.E., Patrinos, G.P.. & Philipsen, S. (2010) Haploinsufficiency for the erythroid transcription factor KLF1 causes hereditary persistence of fetal hemoglobin. Nat Genet, Vol.42, No.9, (September 2010), pp. 801-805, ISSN 1061-4036

Brunkow, M.E., Gardner, J.C., Van Ness, J., Paeper, B.W., Kovacevich, B.R., Proll, S., Skonier, J.E., Zhao, L., Sabo, P.J., Fu, Y-H., Alisch, R.S., Gillett, L., Colbert, T., Tacconi, P., Galas, D., Hamersma, H., Beighton, P., & Mulligan, J.T. (2001) Bone dysplasia sclerosteosis results from loss of the SOST gene product, a novel cystine knot-containing protein. Am J Hum Genet, Vol.68, No.3, (March 2001), pp. 577-589, ISSN 0002-9297

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., & Werner, T. (2005) MatInspector and beyond: promoter analysis based on transcriptional factor binding sites. Bioinformatics, Vol.21, No.13, (July 2005), pp. 2933 – 2942, ISSN 1367-4803

Cartegni, L., Chew, S.L., & Krainer A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet, Vol.3, No.4, (April 2002), pp. 285-298, ISSN 1471-0056

Cassar, M. (2008) Linkage analysis in a familial case of idiopathic epilepsy and its implications in drug development. PhD Dissertation, University of Malta

Farrugia, R., Scerri, C.A., Attard Montalto, S., Parascandalo, R., Neville, B.G.R., & Felice, A.E. (2007) Molecular genetics of the tetrahydrobiopterin (BH4) deficiency in the Maltese population. Mol Genet Metab, Vol.90, No.3, (March 2007), pp. 277 – 283, ISSN 1096-7192

Gianfrancesco, F., Esposito, T., Ombra, M.N., Forabosco, P., Maninchedda, G., Fattorini, M., Casula, S., Vaccargiu, S., Casu, G., Cardia, F., Deiana, I., Melis, P., Falchi, M., & Pirastu, M. (2003) Identification of a novel gene and a common variant associated with uric acid nephrolithiasis in a Sardinian genetic isolate. Am J Hum Genet, Vol.72, No.6, (June 2003), pp. 1479 – 1491, ISSN 0002-9297

Gong Y, Slee, R.B, Fukai, N., Rawadi, G., Roman-Roman, S., Reginato, A.M., Wang, H., Cundy, T., Glorieux, F.H., Lev, D., Zacharin, M., Oexle, K., Marcelino, J., Suwairi, W., Heeger, S., Sabatakos, G., Apte, S., Adkins, W.N., Allgrove, J., Arslan-Kirchner, M., Batch, J.A., Beighton, P., Black, G.C., Boles, R.G., Boon, L.M., Borrone, C., Brunner, H.G., Carle, G.F., Dallapiccola, B., De Paepe, A., Floege, B., Halfhide, M.L., Hall, B., Hennekam, R.C., Hirose, T., Jans, A., Jüppner, H., Kim, C.A., Keppler-Noreuil, K., Kohlschuetter, A., LaCombe, D., Lambert, M., Lemyre, E., Letteboer, T., Peltonen, L., Ramesar, R.S., Romanengo, M., Somer, H., Steichen-Gersdorf, E., Steinmann, B., Sullivan, B., Superti-Furga, A., Swoboda, W., van den Boogaard, M.J., Van Hul, W., Vikkula, M., Votruba, M., Zabel, B., Garcia, T., Baron, R., Olsen, B.R., & Warman, M.L; Osteoporosis-Pseudoglioma Syndrome Collaborative Group. (2001) LDL receptor-related protein 5 (LRP5) affects bone accrual and eye development. Cell, Vol.107, No.7, (July 2001), pp. 513 – 523, ISSN 0092-8674

Hanson, R.L., Ehm, M.G., Pettitt, D.J., Prochazka, M., Thompson, D.B., Timberlake, B., Foroud, T., Kobes, S., Baier, L., Burns, D.K., Almasy, L., Blangero, J., Garvey, W.T., Bennett, P.H., & Knowler, W.C. (1998) An autosomal genomic scan for loci linked with type II diabetes mellitus and bone-mass index in Pima Indians. Am J Hum Genet, Vol.63, No.4, (October 1998), pp. 1130 – 1138, ISSN 0002-9297

Hayer, S., Steiner, G., Görtz, B., Reiter, E., Tohidast-Akrad, M., Amling, M., Hoffmann, O., Redlich, K., Zwerina, J., Skriner, K., Hilberg, F., Wagner, E.F., Smolen, J.S., & Schett, G. (2005) CD44 is a determinant of inflammatory bone loss. J Exp Med, Vol.201, No.6, (March 2005), pp. 903 – 914, ISSN 0040-8724

Heiman, G.A. (2005) Robustness of case-control studies to population stratification. Cancer Epidemiol Biomarkers Prev, Vol.14, No.6, (June 2005), pp. 1579 – 1582, ISSN 1055-9465

Kambouris, M. (2005) Target gene discovery in extended families with type 2 diabetes mellitus. Atheroscler Suppl, Vol.6, No.2, (May 2005), pp. 31 – 36, ISSN 1567-5688

Kania, J.R., Kehat-Stadler, T., & Kupfer, S.R. (1997) CD44 antibodies inhibit osteoclast formation. J Bone Miner Res. Vol.12, No.8, (August 1997), pp. 1155 – 1164, ISSN 0884-0431

Khan, A.A., Bachrach, L., Brown, J.P., Hanley, D.A., Josse, R.G., Kendler, D.L., Leib, E.S., Lentle, B.C., Leslie, W.D., Lewiecki, E.M., Miller, P.D., Nicholson, R.L., O'Brien, C., Olszynski, W.P., Theriault, M.Y., & Watts, N.B. (2004) Standards and guidelines for performing central dual-energy X-ray absorptiometry in premenopausal women, men, and children. J Clin Densitom, Vol.7, No.1, (Spring 2004), pp. 51 – 64, ISSN 1094-6950

Kong, A., & Cox, N.J. (1997) Allele sharing models: LOD scores and accurate linkage tests. Am J Hum Genet, Vol.61, No.5, (November 1997), pp. 1179 – 1188, ISSN 0002-9297

Koziell, A., Grech, V., Hussain, S., Lee, G., Lenkkeri, U., Tryggvason, K., & Scambler, P. (2002) Genotype/phenotype correlations of NPHS1 and NPHS2 mutations in nephrotic syndrome advocate a functional inter-relationship in glomerular filtration. Hum Mol Genet, Vol.11, No.4, (February 2002), pp. 379 – 388, ISSN 0964-6906

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., & Lander, E. (1996) Parametric and non-parametric linkage analysis, a unified multipoint approach. Am J Hum Genet, Vol.58, No.6, (June 1996), pp. 1347 – 1363, ISSN 0002-9297

Lander, E., & Kruglyak, L. (1995) Genetic dissection of complex traits, guidelines for interpreting and reporting linkage results. Nat Genet, Vol.11, No.3, (November 1995), pp. 241 – 247, ISSN 1061-4036

Lindner, T.H., & Hoffmann, K. (2005) EasyLINKAGE: a PERL script for easy and automated two-/multipoint linkage analyses. Bioinformatics, Vol.21, No.3, (February 2005), pp. 405 – 407, ISSN 1367-4803

Markianos, K., Daly, M.J., & Kruglyak, L. (2001) Efficient multipoint linkage analysis through reduction of inheritance space. Am J Hum Genet, Vol.68, No.4, (April 2001), pp. 963 – 977, ISSN 0002-9297

Nyholt, D.R. (2002) GENEHUNTER, Your 'one-stop shop' for statistical genetic analysis? Hum Hered, Vol.53, No.1, (March 2002), pp. 2 – 7, ISSN 0001-5652

O'Connell, J.R., & Weeks, D.E. (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet, Vol.63, No.1, (July 1998), pp. 259 – 266, ISSN 0002-9297

Ott, J. (1991) Analysis of human genetic linkage. John Hopkins University Press, ISBN 0-801-842573, USA

Peltonen, L. (2000) Positional cloning of disease genes: Advantages of genetic isolates. Hum Hered, Vol.50, No.1, (January 2000), pp. 65 – 75, ISSN 0001-5652

Rozen, S., & Skaletsky, H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In: Bioinformatics Methods and Protocols: Methods in Molecular Biology, S. Krawetz, & S. Misener, (Ed.), 365-386, Humana Press, ISBN 0896037320,Totowa, NJ, USA

Seelow, D., Schwarz, J.M., & Schuelke, M. (2008) GeneDistiller--distilling candidate genes from linkage intervals. PLoS One. Vol.3, No.12, (December 2008), e3874, ESSN 1932-6203

Strauch, K., Fimmers, R., Baur, M.P., & Wienker, T.F. (2003) How to model a complex trait. Hum Hered, Vol.55, No.4, (October 2003), pp. 202 – 210, ISSN 0001-5652

Strauch, K., Furst, R., Ruschendorf, F., Windemuth, C., Dietter, J., Flaquer, A., Baur, M.P., & Wienker, T.F. (2005) Linkage analysis of alcohol dependence using MOD scores. BMC Genet, Vol.6, Suppl.1, (December 2005), S162, ESSN 1471-2156

Styrkarsdottir, U., Cazier, J.B., Kong, A., Rolfsson, O., Larsen, H., Bjarnadottir, E., Johannsdottir, V.D., Sigurdadottir, M.S., Bagger, Y., Christiansen, C., Reynisdottir, I., Grant, S.F.A., Jonasson, K., Frigge, M.L., Gulcher, J.R., Sigurdsson, G., & Stefansson, K. (2003) Linkage of osteoporosis to chromosome 20p12 and association to BMP2. PLoS Biol, Vol.1, No.3, (November 2003), pp. 351 – 360, ISSN 1544 - 9173

Thiele, H., & Nürnberg, P. (2004) HaploPainter: a tool for drawing pedigrees with complex haplotypes. Bioinformatics. Vol.21, No.8, (April 2004), pp. 1730-1732, ISSN 1367-4803

Vidal, C., Borg, J., Xuereb-Anastasi, A., & Scerri, C.A. (2009a) Variants within protectin (CD59) and CD44 genes linked to an inherited haplotype in a family with coeliac disease. Tissue Antigens, Vol.73, No.3, (March 2009), pp. 225 – 235, ISSN 0001-2815

Vidal, C., Cachia, A., & Xuereb-Anastasi, A. (2009) Effects of a synonymous variant in exon 9 of the CD44 gene on pre-mRNA splicing in a family with osteoporosis. Bone Vol.45, No.4, (October 2009), pp. 736 – 742, ISSN 8756-3282

Vidal, C., Galea, R., Brincat, M., & Xuereb-Anastasi, A. (2007) Linkage to chromosome 11p12 in two Maltese families with a highly penetrant form of osteoporosis. Eur J Hum Genet, Vol.15, No.3, (March 2007), pp. 800 – 809, ISSN 1018-4813

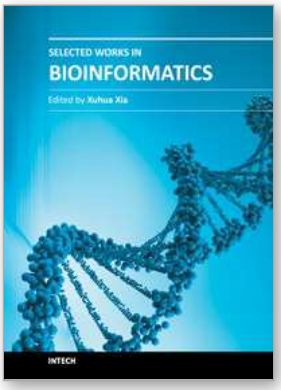
Wang, Z., Rolish, M. E., Yeo, G., Tung, V., Mawson, M. & Burge, C. B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell Vol.119, No.6, (December 2004), pp. 831-845, ISSN 0092-8674

Wright, A., Charlesworth, B., Rudan, I., Carothers, A., & Campbell, H. (2003). A polygenic basis for late-onset disease. Trends Genet, Vol.19, No.2, (February 2003), pp. 97 – 106, ISSN 0168-9525

Wright, A.F., Carothers, A.D., & Pirastu, M. (1999) Population choice in mapping genes for complex diseases. Nat Genet, Vol.23, No.4, (December 1999), pp. 397 – 404, ISSN 1061-4036

Xu, J., & Meyers, D.A. (1998) Lod Score Analysis, In : Approaches to Gene Mapping in Complex Human Diseases, Haines, J.L., & Pericak-Vance, M.A., pp. 253 – 272, Wiley-Liss Inc., ISBN 0-471-17195-6, USA

Zlotogora, J. (2007) Multiple mutations responsible for frequent genetic diseases in isolated populations. Eur J Hum Genet, Vol.15, No.1, (January 2007), pp. 272 – 278, ISSN 1018-4813



Selected Works in Bioinformatics

Edited by Dr. Xuhua Xia

ISBN 978-953-307-281-4
Hard cover, 176 pages
Publisher InTech
Published online 19, October, 2011
Published in print edition October, 2011

This book consists of nine chapters covering a variety of bioinformatics subjects, ranging from database resources for protein allergens, unravelling genetic determinants of complex disorders, characterization and prediction of regulatory motifs, computational methods for identifying the best classifiers and key disease genes in large-scale transcriptomic and proteomic experiments, functional characterization of inherently unfolded proteins/regions, protein interaction networks and flexible protein-protein docking. The computational algorithms are in general presented in a way that is accessible to advanced undergraduate students, graduate students and researchers in molecular biology and genetics. The book should also serve as stepping stones for mathematicians, biostatisticians, and computational scientists to cross their academic boundaries into the dynamic and ever-expanding field of bioinformatics.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821