Center for Genomics and Bioinformatics
Karolinska Institute
Stockholm, Sweden

SNP TECHNOLOGY AND ALZHEIMER'S DISEASE

Walter Mathias Howell

Stockholm 2003

Published and printed by Karolinska University Press
Box 200, SE-171 77 Stockholm, Sweden

ISBN 91-7349-473-9

To the lights in my life, Åsa & Liv

ABSTRACT

One major goal of genetic research is to understand the role of genetic variation. By far the most common type of such variation in humans involves single DNA bases, and is termed single nucleotide polymorphism (SNP). With sufficient technological solutions, one strong belief is that SNPs can enable the mapping of disease genes involved in complex genetic disorders.

Alzheimer's disease (AD) is a complex disorder characterized by progressive cognitive decline and memory impairment. Some individuals acquire this form of dementia before the age of 65 (referred to as early-onset or familial AD) but most often AD occurs late in life. It is in the early-onset form, however, where causative mutations have been found in three different genes; APP, PSEN1, & PSEN2. Other than these, the only additional risk factor identified for AD is the ε4 allele of the APOE gene. Together these only account for a fraction of AD, leaving room for studies to identify additional AD susceptibility genes.

In the initial investigation of this thesis, polymorphisms in the PSEN1, PSEN2, APOE and VLDL-R genes were tested for association with early-onset Alzheimer's disease (EOAD). Aside from confirming the well-established APOE-ε4 association, an allele of the PSEN2 showed a significant disease association. In an attempt to verify the PSEN2 association and localize the potential pathogenic variant, a series of eight additional SNPs, located throughout the PSEN2 gene, were tested for association with EOAD. None of the tested markers showed significant disease association upon replication in a second set of cases and controls.

The remainder of the thesis describes the invention and development of a novel technique for scoring SNPs. The method, called Dynamic Allele Specific Hybridization (DASH), is a great improvement in both throughput and reliability compared to traditional techniques. The crucial step in the procedure is the heating of the DNA duplex (formed from hybridization of the PCR-amplified target with an allele-specific probe) whilst measuring the fluorescence of a double-strand specific intercalating dye. SNP alleles are detected and scored by comparative analysis of the melting profiles.

Improvements to the initial DASH format are detailed in the final two papers of this thesis. A novel alternative in fluorescence detection, termed *induced* fluorescent resonance energy transfer (iFRET), is introduced. IFRET employs energy transfer between a generic intercalating dye and an FRET acceptor attached to the allele-specific probe. This system retains the spectral-multiplex potential offered by traditional FRET systems, while reducing costs and improving fluorescence signal intensities. Aside from detection, DASH was converted from a microtiter-plate format to an array format which greatly improved flexibility and simplified the assay procedure. The complete DASH-2 system is examined in terms of multiplex options, throughput, cost and accuracy.

LIST OF PUBLICATIONS

This thesis is based on the following articles, which are referred to by their roman numerals:

I Brookes AJ, **Howell WM**, Woodburn K, Johnstone EC, Carothers A.
Presenilin-I, presenilin-II, and VLDL-R associations in early onset Alzheimer's disease.
Lancet (1997) Aug2;350(9074):336-337

II **Howell WM**, Brookes AJ.
Evaluation of multiple presenilin 2 SNPs for association with early-onset sporadic Alzheimer disease.
Am J Med Genet (2002) 111;2:157-163

III **Howell WM**, Jobs M, Gyllensten M, Brookes AJ.
Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms
Nat Biotechnol. (1999) Jan;17(1):87-8.

IV **Howell WM**, Jobs M, Brookes AJ.
iFRET: An Improved Fluorescence System for DNA-Melting Analysis
Genome Res. (2002) 12:1401-1407.

V Jobs M*, **Howell WM***, Strömqvist L, Brookes AJ.
DASH-2: Flexible, low-cost and high-throughput SNP genotyping by Dynamic Allele Specific Hybridization on membrane arrays.
Manuscript.

*These authors contributed equally to this work.

CONTENTS

LIST OF ABBREVIATIONS

AFM	Atomic force microscopy
APEX	Array primer extension
APOE	Apolipoprotein E
APP	Amyloid precursor protein
Aβ	Amyloid-β protein
Bp	Base-pair(s)
χ^2	Chi-square
CCD	Charged-couple device
CDC	Center for disease control
CDCV	Common disease common variant (theory)
CF	Cystic fibrosis
cSNP	Coding single nucleotide polymorphism
DASH	Dynamic allele specific hybridization
DGGE	Denaturing gradient gel electrophoresis
DNA	Deoxy-ribonucleic acid
DZ	Dizygotic
EOAD	Early-onset Alzheimer's disease
FAD	Familial Alzheimer's disease
FEN	Flap endonuclease
FP	Fluorescence polarization
FRET	Fluorescence resonance energy transfer
HGP	Human genome project
HGVBase	Human genome variation database
htSNP	Haplotype tagging single nucleotide polymorphism
iFRET	*induced* fluorescence resonance energy transfer
Kb	10^3 base-pairs
LD	Linkage disequilibrium
LOAD	Late-onset Alzheimer's disease
MALDI-TOF MS	Matrix assisted laser desorption/ionization time-of-flight mass spectroscopy
Mb	10^6 base-pairs
MZ	Monozygotic
OLA	Oligo ligation assay
OMIM	Online Mendelian inheritance in man
OR	Odds ratio
PCR	Polymerase chain reaction
PHF	Paired helical filaments
PSEN1	Presenilin 1
PSEN2	Presenilin 2
QY	Quantum yield
RCA	Rolling circle amplification
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RR	Relative risk
SBE	Single base extension
SNP	Single nucleotide polymorphism
SSCP	Single stranded conformation polymorphism
STEM	Scanning transmission electron microscopy
Tm	Melting temperature
TSC	The SNP consortium
VNTR	Variable number tandem repeats
ε	Extinction coefficient
λ_R	Familial clustering

INTRODUCTION

Of all the scientific advances during the last century, the production of a working draft of the human genome is probably one of the most publicly noted accomplishments. This DNA sequence, composed of over 3 billion base-pairs (bp), identifies what people have in common. The current challenge is to discover what genetic differences exist between individuals, and determine their impact on human health.

The study of variation in the human genome forms the core of this thesis. One major line of research involves the use of genetic polymorphism for the dissection of multi-factorial disease. Another is the invention, development, and implementation of technology that is necessary for efficient use of genetic variation.

In order to place this work into the context of the modern field of genetic research, I will begin with a brief description of genetic variation, followed by a discussion of genetic approaches to solve complex disease, and finish with an overview of methods and technologies involved in DNA-variation analysis.

Throughout this thesis, important concepts will be presented in callouts. Short definitions or examples will be given to illustrate the point and to further enhance reader understanding.

DNA:

Deoxy-ribonucleic Acid – the chemical material of inheritance for living organisms.

GENETIC VARIATION

The genetic material inside all living organisms is made from the same basic components, namely the nucleotides Adenosine, Guanosine, Cytidine, and Thymidine. These four building blocks are linked together into long chains, the sequence of which then "codes" for the various proteins and gene products that the organism needs to survive. The precise collection and organization of these DNA sequences is specific for each species, and is called a *genome*.

However, just as no two snowflakes are exactly alike, rarely do even members of the same species have the exact same genetic makeup (the notable exception being identical twins). On average, any two humans share a 99% genetic identity [1]. Although the majority of the differences in DNA sequence (genotype) do not result in any noticeable physical change (phenotype), the few that do account largely for the diversity we see in the human population as far as height, eye, skin, and hair color, etc.

From mutation to polymorphism

For the average human being, the process of growing from a single fertilized egg to an 60 year old adult involves something in the order of 1,000,000,000,000,000,000 (10^{17}) cell divisions [2]. Every cell division requires the copying (replication) of the entire 3 billion bp sequence of the human genome. Sophisticated proof-reading and repair mechanisms [3] do their best to maintain fidelity, but replication errors do occur. The process of creating new genetic variation is called "*mutation"*. On average, the chance that any one particular base will undergo a mutation is in the order of 10^{-8} per cell generation [4]. Mutations can arise from natural internal processes such as cell replication, meiotic recombination, gene conversions, and also from any number of environmental factors such as radiation or free radical damage caused by ingestion of toxins.

In humans, every individual has two copies of the genome, one originating from each parent. So at a given location (*locus*) or position in the genome, each individual has two copies of the particular sequence. When a mutation occurs, the event causes a change in one DNA sequence resulting in the individual having one copy of the original sequence and second new sequence at the mutation locus. The two alternative versions of DNA sequence at the mutation locus are referred to as *alleles*. (figure 1)

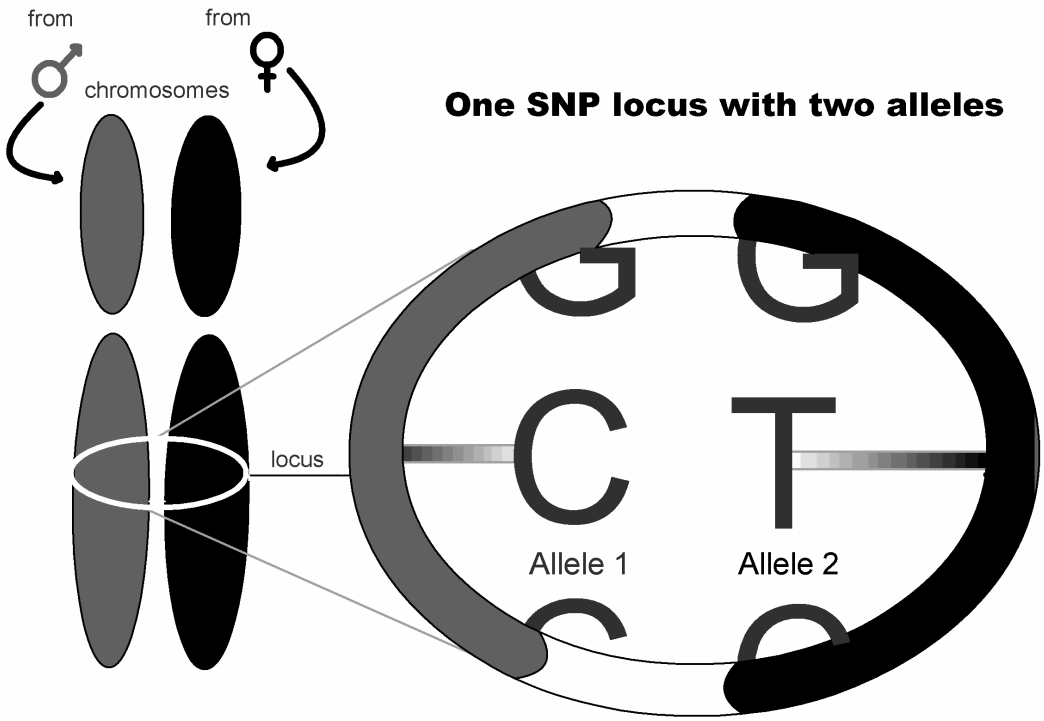


Figure 1: A locus is a physical location on a chromosome. Alleles are alternative DNA sequences that can be found at a locus.

Most mutations arise in differentiated *"somatic"* cells, and remain only in the individual in which the mutations occurred. However, if the mutations occur in *"germ cells"* (egg or sperm cells), then these genetic changes can be passed on to one's offspring and are thus inheritable.

The fate of an inherited mutation, or to be more specific, the fate of the newly created *allele* is determined by a number of factors. If the mutation is deleterious (reduces fitness) then *negative selection* is likely to quickly act to reduce or remove the mutant allele from the *gene pool*. If on the other hand, the mutation is advantageous, then *positive selection* will likely work to increase the frequency of the new allele in the population. As stated earlier, however, the vast majority of mutations cause genetic changes that are neutral or have very mild effects. The third force is thus the random chance that the mutation will pass on to successive generations. This phenomenon, termed *genetic drift*, can act to either increase or decrease allele frequencies in the population. Regardless of the means by which it occurs, if the mutant allele reaches a frequency of 1% or more in the population, then the locus is said to be *polymorphic*.

Gene Pool:

The sum of all genetic variation in a population.

Before continuing to describe the numerous forms of genetic polymorphism, I would like to stress that mutation is a normal and essential process for any organism. The differences in DNA sequence generated from mutational events give rise to the vast amounts diversity we observe, and in the long run assist in an organism's ability to adapt to new environments. In every day life, the term mutation often evokes an image of a three-headed monster or some strange superhero with superhuman powers. Indeed mutagenic compounds, such as found in cigarette smoke, can give rise to various maladies including cancer. However, not all mutations are bad, and we all have a constellation of thousands of mutations that form our individuality.

Genetic variation:

A position in the genome where the DNA sequence can vary between individuals.

Polymorphism types

"Polymorphisms add colors to the palette, allowing us to paint a more refined picture of mankind and his relation to his genome"

WMH

The word "polymorphism" comes from Greek and means "having many shapes". When describing genetic polymorphism, it is easiest to imagine two strands of DNA that differ in sequence rather than absolute shape. In this way, the most common types of polymorphism found in the human genome can be organized into the three classes or categories.

Repetitive Elements:

Certain DNA sequences are found in multiple copies throughout the genome. One subclass consists of relatively long sequences that are spread fairly evenly among chromosomes. A classic example of which is the ALU repeat that is around 330 bp in length, and found in over 750,000 copies in the genome [5]. Another form of repeat polymorphism is referred to as Simple Tandem Repeat Polymorphisms (STRPs) or "*microsatellites*" [6, 7]. This type of polymorphism consists of short di-, tri-, or tetra-nucleotide units that are consecutively repeated at the polymorphic position (figure 2). The number of repeat units varies between alleles. Microsatellites are often highly polymorphic, having up to 30 alleles [8], and thus show a high *allelic diversity* and show high *heterozygosity*.

Heterozygosity:

The likelihood that any two randomly chosen chromosomes will have two different alleles at the polymorphic site.

Insertion and Deletions:

Insertion-deletion or *indels* are di-allelic polymorphisms. The difference between alleles lies in the presence or absence of one or more DNA bases at the polymorphic position. (figure 2).

Substitutions:

Substitutions are also most often di-allelic polymorphisms. Alleles of this type are distinguished by replacement of DNA bases, rather than presence or absence as in indels. (figure 2).

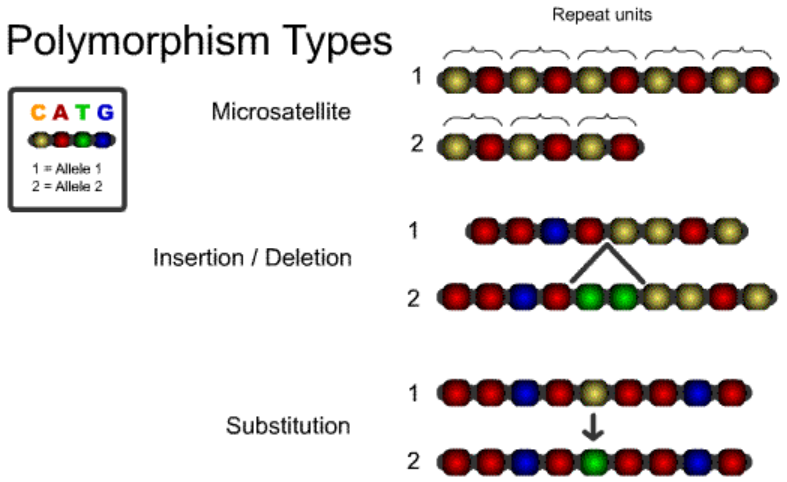


Figure 2: Three different classes of polymorphism commonly found in the human genome. Note*only a single strand of DNA is represented for each allele.

Single Nucleotide Polymorphisms

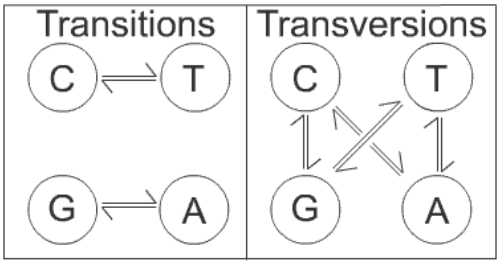
The key defining character of *Single Nucleotide Polymorphisms* (SNPs) is that alleles of these polymorphisms involve only single bases. I have created a special section for this type of polymorphism because it is one of the most widely researched and debated forms of polymorphism in contemporary genetic research. In addition, every publication in this thesis, in one way or another, involves the study of SNPs.

Strictly speaking, SNP alleles can exist as insertion or deletion of a single base, or substitution of one base for another. In the case of substitution, the maximum number of SNP alleles is limited to just four. This is because DNA is made up of only 4 different nucleotide bases (abbreviated C, A, T, G) thus single nucleotide substitutions are at most *tetra-allelic*. Tetra-allelic or even tri-allelic SNPs, however, are very infrequent with the majority of true documented cases being in the mitochondrial genome [9]. For this reason, SNPs are generally thought of as di-allelic polymorphisms. Also, interpretation of the exact definition of an SNP leads some individuals to consider only substitutions rather than indels as SNPs. For simplicity and consistency throughout the remainder of this thesis, SNPs will be considered as di-allelic, single nucleotide substitutions.

Single Nucleotide Polymorphism:

Single base pair positions in genomic DNA where; a) different sequence alternatives exist in normal individuals in some population, and b) the least frequent sequence alternative is greater than 1%.

SNP alleles are created either by *transition* (purine-purine, or pyrimidine-pyrimidine substitution) or *tranversion* (purine-pyrimidine or pyrimidine-purine substitution) [10]. All of these transitions and transversions events appear to be more or less similar in occurrence, except for the extreme overabundance of the C to T transitions. Over 70% of all SNPs found in the human genome involve a C to T transition. This is likely due to the chemical conversion of 5-methyl Cytosine residues to Thymidine through a deamination mechanism [11]. Coincidently, this process should lead to humans having a slightly more A-T rich genome with each successive generation, but I have not been able to find discussion of this topic in the published literature.



We now know that the large majority of all polymorphic positions in the genome are SNPs. However, even as recently as April of 1999, there were only some 7,000 SNPs available in the public domain [12]. Less than 2 years later, in a single publication, an additional 1.42 million predicted SNPs were released into the scientific community [13]. The work was performed as a combined effort between the academic Public Human Genome Project (HGP) [14], and a joint academic/industrial organization called The SNP Consortium (TSC) [15]. At the time of writing this thesis, the number of available SNPs has grown to over 3 million. These SNPs can be found in either of the two major repositories for SNPs, namely dbSNP at (www.ncbi.nlm.nih.gov/SNP/), and HGVbase currently located at (hgvbase.cgb.ki.se). Both databases are publicly accessible and free to use for the scientific community.

Distribution – chromosomal

Although there are some 3 million SNPs already in the databases, this is only a fraction of the estimated 11 million SNPs thought to reside in the human genome [16]. This estimate would lead to a genome-wide average of 1 SNP per every 300 Kb of sequence in the population. A number of studies have pointed to a genome-wide average of about 1 SNP occurring in 1-2kb of sequence when comparing any two randomly chosen chromosomes [17-22]. In one large study [13], all 22 human autosomes had a fairly consistent average of 1 SNP per 1.9kb of sequence. Conspicuously, the rate of single nucleotide polymorphism in the X and Y chromosomes was roughly half (1/3.7 Kb and 1/5.2 Kb respectively). The decrease of polymorphism in the sex chromosomes may be attributed to the reduced amount of recombination inherent in these chromosomes during meiosis [23].

SNPs, however, are not evenly distributed down the length of any chromosome. The genetic architecture of human chromosomes is often marked by large stretches of non-coding sequences, with patches of gene clusters containing coding sequence. Genetic variation is roughly four times less in coding sequence than in non-coding sequence [20, 24]. It seems reasonable that natural selection pressure would act to preserve certain sequences such as exons, promoters, and enhancer sequences as alteration of these sequences could adversely affect normal biological functions. There are few exceptions, however, where coding sequence show a high degree of polymorphism. For instance, in and around the HLA genes (which encode important components of the immune system) there is high sequence variability [25]. This is thought to be the result

selection pressure to maintain variability and thus allow for recognition of a larger repertoire of antigens.

Many of the SNPs that occur in exons, occur in the wobble position of the reading frame, and thus do not alter the protein sequence. These changes are called *synonymous* or silent substitutions and are thought to have little or no effect on the gene product. *Non-synonymous* variants, on the other hand, cause a substitution of one amino-acid for another at the protein level. The consequence of non-synonymous substitution on protein function varies from no effect to complete disruption of normal protein function. The most severe single base changes in exons can produce detrimental anomalies such as shifts in the open reading frame, or even creation of a stop codon, either of which can cause non-functional copy of the gene product. These severe types of non-synonymous variants rarely if ever reach a high enough frequency to be considered polymorphisms.

Two additional regions of the chromosome, the ends (telomeres) and the center (centromere), have important biological functions and are known to be highly polymorphic. Telomeres play an important role in aging [26], and centromeres are key elements in cell division [27]. These regions are difficult to study, however, because the complexity of repeats makes these regions difficult to sequence, and of the efforts so far, these regions appear to be very gene-poor [26].

All of the numbers given here should be taken with a grain of salt. The task of pinpointing the exact location of SNPs on a chromosomal map has proven problematic. One reason is that the human reference sequence [28], which the SNPs are being mapped to, is just a working draft. The sequence is constantly being updated, rearranged, and augmented, which leaves SNP mapping efforts with a bit of a moving target. Gene families (genes with similar sequence) and pseudogenes tend to complicate matters as the SNP maps to several positions. Also, the SNP databases are known to contain many artifacts and false SNP predictions. In my opinion, the extreme haste to discover and "brand" SNP entries in the databases has resulted in low quality thresholds where sequence misalignments, sequencing errors, annotation errors, and other anomalies are far too common. A recent study suggests that as much as *50%* of the entries in databases are spurious [29]. This being said, the quality of database entries is constantly improving through SNP validation efforts and increased curation of the databases. The availability of "*in silico*" or computer-based searches for SNPs has greatly reduced experimental set-up time and sometimes cost for many experiments.

Distribution – geographic

The recent avalanche of SNP data has provided population geneticists with some new and useful tools to investigate human populations world-wide. Although there have been only a few thorough studies of human polymorphism on the global scale, the estimate is that somewhere between 80-95% of all SNPs can be found in all major population groups [30,31]. A very large percentage, suggesting that most of the SNPs we discover today are very old and that only a fraction (5-20%) of SNPs discovered are population specific [32]. Studies on several different SNPs in multiple populations have shown that allele frequencies can vary quite dramatically. As an example, the C allele of a C/T polymorphism located in intron 2 of the dopamine D2 receptor gene reaches a frequency of 60% in the Irish population, and is lower than 1% in several South American Indian and East Asian popultions (figure 3). By detailed survey of allele frequencies in modern day populations, it is possible to make inferences into human origins and historical migration patterns.

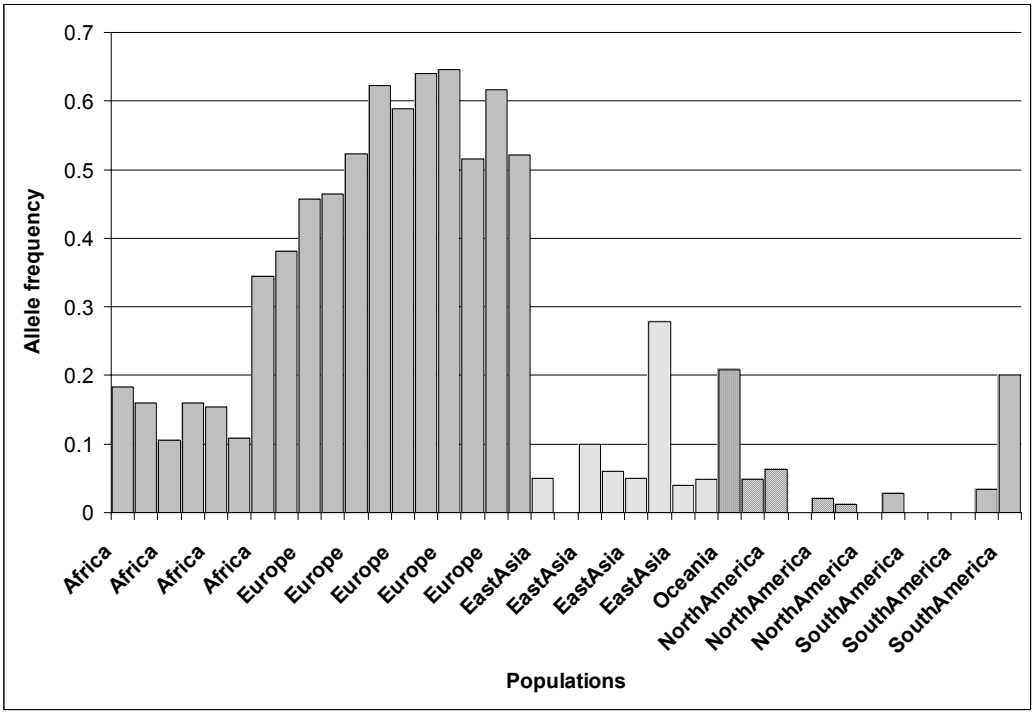


Figure 3: Allele frequencies of Dopamine D2 SNP as reported from different populations around the world. Source of the SNP frequency data is the web-based Allele Frequency Database (ALFRED) accessible through (http://info.med.yale.edu/genetics/kkid)

Navigating the genome

Physical maps, genetic maps, and linkage disequilibrium

Before leaving the basic discussion of polymorphism and genetic variation, several concepts regarding analysis of the relative location of SNPs and other polymorphisms need to be addressed. Just as a traveler in a car benefits from having a map when deciding how to get from one city to another, it is imperative for a genetic researcher to be able to locate the relative location of different features on a chromosome. Therefore, a number of different maps have been created to assist in navigating the genome.

The most basic framework to visualize the human genome is called the *physical map*. The ultimate physical map is the actual base by base sequence of the DNA contained on each chromosome [33]. Humans have a total of 46 chromosomes consisting of 22 pairs of autosomes, and a pair of sex chromosomes (two X chromosomes in females, and one X and one Y in males). The chromosomes are ordered roughly by size, and the shortest chromosomes (21 & 22) are roughly 33 Mb in length [13]. At the time of writing this thesis, it is only these two shortest chromosomes that have high-quality "finished" sequence, and thus have a complete physical map [34, 35]. The sequences of the other chromosomes are in various stages of production, so the exact physical map of these chromosomes is as of yet incomplete.

An additional navigation aide consists of unique reference sequences commonly called *genetic markers*. These are typically some type of polymorphism that are located at specific locations on chromosomes. These sequences usually do not affect a person's health, but can act as genetic "mile-markers" or waypoints to determine relative distances between the marker and a gene, mutation, or other genetic feature. On physical maps, the distance between two genetic markers is measured in base-pairs.

Genetic maker:

A polymorphism or other genetic feature that can be uniquely located on a chromosome.

Another type of map that is commonly used is called a *genetic map*. To understand the distance units used in genetic maps, it is good to review meiosis, the cell divisions that lead to gamete (egg or sperm) formation. One step in this process involves the alignment of pairs of homologous chromosomes and exchange of genetic material through a process called *crossing-over*. Crossover events thus act to shuffle the DNA sequences from paternal and maternal chromosomes just prior to production of the gametes. Several crossover events or *recombinations* can occur on each chromosome per meiosis. Genetic markers that lie close to each other on a chromosome are less

likely to be separated by recombination than markers that are located far apart. On a genetic map, the unit used to describe the distance between two genetic markers is thus a measure of the likelihood that a recombination event will occur between the two markers. The primary unit is called a centiMorgan (cM) and is equivalent to 1% recombination. Recombinations tend to occur more often in the creation of eggs than creation of sperm, so the genetic map is roughly 1.5 larger for females (~4400 cM) than for males (~2700 cM) [36]. Also, on chromosomes there tends to be "hot spots" or "jungles" [33] where recombination happens more frequently•, and "deserts" where recombination is rare. This results in recombination rates that vary at least two orders of magnitude across the genome. For all these reasons, it is impossible to have one accurate conversion factor between genetic and physical maps, but a common estimate is that 1 cM is equivalent to approximately 1 Mb of DNA sequence.

Recombination:

Reciprocal exchange of genetic material between homologous chromosomes

The physical map is thus a representation of the physical string of Gs, As, Ts, and Cs that make up each chromosome, while the genetic map is based on and describes an important biological process. Both types of maps are useful, and are employed in some later sections of this thesis.

One last important concept related to the ideas of recombination and mutation is *linkage disequilibrium* (LD). As mentioned earlier, chromosomes abound with a large variety of polymorphic loci. When a mutation occurs, the newly created sequence variant is surrounded by a number of alleles from neighboring polymorphic loci. If the new variant is passed on through successive generations, those alleles from the nearest neighboring polymorphic loci are likely to follow and be co-inherited on the same physical piece of DNA. Hence, the presence of the new variant makes it possible to predict the identity of the nearby alleles. *Linkage disequilibrium* is thus a measure of the tendency of alleles to occur together on a chromosome. In other words, LD measures the extent to which alleles occur more often on the same chromosome than expected by random segregation [37]. In addition, a collection of alleles found together on a chromosome is called a *haplotype*.

Linkage Disequilibrium:

The non-random assortment of alleles

• SNP variability is higher in regions of high recombination. It is postulated now that the recombination process is mutagenic, giving rise to higher numbers of SNPs in recombination hot spots [23].

The shuffling process of recombination acts to decrease LD and create new combinations of alleles (haplotypes). Recombination in the first few generations is likely to separate alleles from loci that are far apart, however after many generations, even loci located as close as 5 bp can show little or no LD. The extent and fluctuation of LD in a population is dependant on a number of factors besides recombination. In small populations, genetic drift can act to quickly remove or "fix" alleles and haplotypes from the genepool, increasing LD over larger chromosomal regions [38, 39]. To these ends, genetic drift works more slowly in larger populations.

The relationship between physical maps, linkage disequilibrium, and genetic maps all assist in navigating through the genome to find the interesting bits, for example when studying disease.

DISEASE RESEARCH

The quest to uncover the causes of disease and to design cures has been evident in civilization as long as there has been writing. As far back as 1600 BC, the Egyptians demonstrated advanced medical knowledge. The Ebers Papyrus (figure 4), an ancient Egyptian document, described diagnosis and over 700 remedies for various maladies, covering everything from heart surgery to asthma. The cause of disease was often believed to be due to a spirit or the ghost of a dead adversary, so treatment often involved rituals and magic in addition to a number of herbal & mineral remedies.

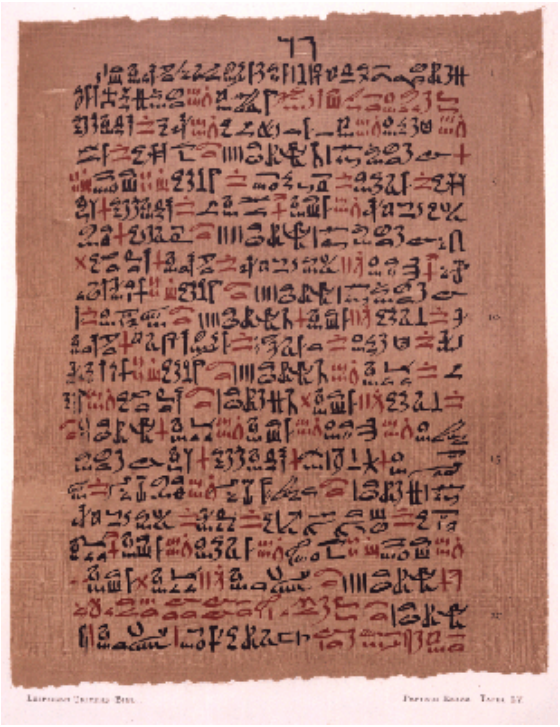


Figure 4: A page from the 4000 year old Ebers Papyrus. The hieroglyphs describe an herbal inhalation therapy against asthma. Picture: courtesy of the National Library of Medicine (USA).

In more recent history, the major cause of fatal disease could be attributed to a number of infectious agents. Millions of people around the globe fell victim to the plague, Spanish flu, or smallpox. But in time our scientific tools, medical theories, and hygiene practices have progressed bringing with it remarkable improvement in human health. In parallel with the smallpox vaccine, the discovery of penicillin (Alexander Fleming, 1929), and its therapeutic implementation as an antibiotic (Howard Walter Florey, 1940), the list of the most common life-threatening diseases in man has changed radically. In 1900, the infectious diseases pneumonia, tuberculosis, and enteritis were the top three leading causes of death accounting for over 30% of all recorded deaths according the Center for Disease Control and Prevention (CDC: www.cdc.gov). Today heart disease and cancer (two afflictions that appear to have a large genetic component) are responsible for half of all deaths in developed nations.

Taming infectious disease resulted in an increase life expectancy to an average of 76 years in western societies in the year 2000, a gain of 30 years since 1900 (CDC: www.cdc.gov). In turn, the medical community of today is facing new problems. An ever aging population is now suffering from a series of diseases that appear late in life,

and if the trends in medicine continues, people will only get older. Discerning the cause of these diseases has proven challenging, and at present the medical field is turning towards genetics for the missing answers. The hope being that unlike the infectious diseases of the past, that the common diseases of today have their roots in discrete genetic defects. Further that identification will lead to cures and therapies. The first question one should ask before embarking on a search for genetic determinants to a specific disease is:

How do we tell if a disease is genetic?

In current research there are several means of evaluating whether or not there is an inherited component to a specific disease. A casual method is to examine if the disease appears to run in families. For example, if the disease is rather rare but there tends to be families that have several affected family members, then suspicion rises for an inherited component. One way is to quantify such a tendency is to measure *familial clustering* (λ_R) [40]. The calculation involves measuring the incidence of the disease for relatives of different degrees (parents, siblings, uncles, etc…) and to divide those numbers by the prevalence of the disease in the population. A large λ_R value indicates strong evidence for a hereditary component.

A cautionary note is that a high λ_R may overestimate the contribution of genetic components. Beyond genetics, relatives also inherit eating habits, active or sedentary lifestyles, and languages just from being reared together. Many of these "*shared environmental*" factors, typically classified as anything not genetic, can have a bearing on disease.

Thus, more precise measures of heredity have been developed to distinguish the genetic components from the shared environment. Twin studies are one such tool [41]. Monozygotic (MZ) twins are genetically identical, and are thus expected to be *concordant*, meaning sharing the same disease state, if the disease is largely genetic. Dizygotic (DZ) twins share 50% of their genetic make-up, just as any other sibling would. DZ twins would be expected to have a lower concordance rate. If there is no difference between the rate of concordance between MZ and DZ twins, yet there is a high λ_R value, then the hereditary culprit is probably the shared-environment. Analysis of twins reared together vs. twins reared apart (adopted) allows for more precise measures of the genetic and shared-environmental components in disease.

Inherited disease can be put into two rather broad categories: Simple diseases that tend to have single genes responsible for disease, and complex diseases that appear to have a genetic component, but no single causative genes can be identified.

Simple disease

The understanding of simple or monogenetic disease really has its roots in the work of an inquisitive monk named Gregor Mendel. In the mid 1860's, he made some key discoveries about inheritance, not by the study of disease but rather some peculiarities with pea plants. By cross-breeding pea plants with different traits like pea shape or flower color, he was able to deduce that some traits are inherited in predictable patterns. Through careful analysis, Mendel reasoned that traits are inherited through independent hereditary units, that one version of each hereditary unit comes from each parent, and that a trait may not show up in an individual but can still be passed on to the next generation. Today Mendel's hereditary unit is called a *gene*, and the different traits or *phenotypes* are caused by inheritance of different *alleles* of a gene. If an individual requires two copies of the same allele for the phenotype to be expressed, the trait is said to be *recessive*. If, on the other hand, the presence of one allele is sufficient for the phenotype to be expressed (regardless if the other allele is the same or not), then the trait is said to be *dominant*.

Over 4,000 different diseases that exhibit some form of mendelian inheritance have been described and deposited in the online database called the Online Mendelian Inheritance in Man (OMIM: http://www.ncbi.nlm.nih.gov/omim/). These diseases are often rare and highly *penetrant*, meaning individuals that have the disease-associated allele are also very likely to also show the disease. When several different alleles or mutations in the same gene can give rise to disease, the disease is said to show *allelic heterogeneity*. A trademark example of a Mendelian disease that shows high allelic heterogeneity is Cystic Fibrosis (CF). Although a single allele accounts for over 70% of all cases of disease, over 1000 rare mutations in the same gene have been reported (Cystic Fibrosis Mutation Database: www.sickkids.on.ca/cftr/).

Complex disease

Any disease that does not follow the basic Mendelian laws of inheritance is considered a complex disease [42]. Instead of following the formula of one gene - one illness, these diseases appear to arise from the interplay between both genetic and non-genetic factors. By convention, if environmental factors are involved, the condition is referred to as *multi-factorial inheritance*. If the interaction between alleles of multiple genes is required to give rise to the disease phenotype, the condition is traditionally called *polygenic inheritance*. If different or overlapping sets of genes are important for disease expression in different individuals, *locus heterogeneity* is at work. If there are many sets disease predisposing alleles in these genes, then *allelic heterogeneity* is at work. In any case, no single environmental or genetic factor appears to be necessary or sufficient to cause disease. Complex disease is thus a collection of risk modifying factors that in summation result in the disease phenotype. Risk alleles are thus more probabilistic than deterministic in constitution [43].

Many common diseases fit well into this category. Cardio-vascular, cancer, and diabetes, all of which are among the top ten leading causes of death (CDC: www.cdc.gov), are considered complex diseases. Several diseases that occur in the elderly, such as Alzheimer's disease and arthrosclerosis, also fall in this category. Although pedigree analysis often yields no distinct patterns of inheritance, heritability studies for at least some of these diseases implicate high genetic contributions. A case in point is late-onset Alzheimer's disease (LOAD) where as much as 74% of the risk can be attributed genetic rather than environmental factors [44].

The search for the genetic determinants of complex disease is a formidable task. Locating or *mapping* the individual disease predisposing alleles in the genome relies on careful study design, accurate genetic measurements, and some would say, a great deal of luck [45]. Although there is no recipe for how to succeed with every complex phenotype, a number of study design considerations can be addressed in order to maximize the chances of success.

SIX FUNDAMENTAL, RECURRENT, CRITICAL NAGGING QUESTIONS

"To know what to ask, is already to know half"

Aristotle

As with most things, success in mapping disease genes pivots on thorough preparation. Most of the following questions and answers revolve around how to set-up an effective search for the genetic component(s) of complex disease. This is not a cookbook however as many of the questions have only partial answers and the answers given are largely a matter of debate. In an effort to untangle the genetic contribution to disease, a few basic questions that often arise are:

1) *Which of the general strategies best fits my genetic disorder?*

There are two common approaches taken in attempts to isolate human disease genes. The fundamental difference between the two strategies lies in the amount of prior knowledge needed about the disease. In the *functional cloning* approach, information about the structure or function of the protein is used to isolate the corresponding gene. *Positional cloning*, on the other hand, requires no previous knowledge of the biochemical or biological basis for the disease. Instead, the strategy involves first identifying the chromosomal region(s) likely to harbor the disease gene(s) followed by isolation of the gene(s) at fault.

The two tools that are vital for these approaches are *linkage studies* and *association studies.* Both of them can make use of neutral polymorphisms (*genetic markers*) to hunt down the mutations or pathogenic changes that are causative for disease.

Linkage studies

Basically, the study of *linkage* tries to see if two factors tend to be inherited together in families. Before DNA was understood, the analysis involved the co-inheritance of two clearly visible phenotypes. An early example was when Bateson and Punnett (1906) were cross-breeding different strains of sweet peas; they noticed that plants with purple flowers were almost exclusively found to have long pollen, whereas red flowers always had round pollen. Only very infrequently did crosses give rise to purple flower-round pollen or red flower-long pollen plants. We know now that this was because the genes for flower petal color and pollen shape are located closely together on the same

chromosome. Cross-over events are thus unlikely to occur between the two loci and disrupt the haplotypes bearing the paired traits. Hence, these genes (and phenotypes) are *linked*.

Rather than study phenotype-phenotype relationships, linkage analysis for mapping disease genes in humans typically involves evaluating genotype-phenotype relationships. Specifically, linkage studies track the inheritance of genetic marker alleles (genotypes) through the pedigree of a family, or set of families, that exhibit a specific disease (phenotype). The closer the marker locus is to the pathogenic locus underlying the disease phenotype, the less frequent crossover events will occur between alleles of the marker and disease locus. In simple terms, a successful linkage study uses genetic markers to tag region(s) of the genome that are likely to harbor the genetic variants responsible for the disease.

Linkage studies:

Measures the recombination events at marker loci in pedigrees to identify chromosomal segments that tend to co-segregate with the disease phenotype.

The process starts with the enrollment of families that show a history of the disease in question and collection of DNA from the individual family members. A *pedigree* is simply a schematic drawing of family relationships and disease status. Figure 5 demonstrates a pedigree consisting of three generations. Squares depict males and circles are females. People having the disease are indicated with filled shapes, and non-affected individuals have hollow shapes. A diagonal line through the square or circle indicates the individual is deceased. Examination of disease status within the pedigree can give clues as to the chromosomal location of the disease causing genetic element. The inheritance pattern in figure 5 is consistent with an autosomal dominant disease model where the gene responsible for the disease (in this family) is located somewhere on chromosomes 1-22 (thus excluding the sex chromosomes). In complex disorders the inheritance pattern is often less obvious.

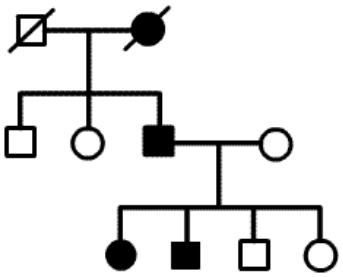


Figure 5: A three generation pedigree where the disease shows autosomal dominant inheritance.

To perform a linkage study over the entire genome (a *genome scan*), a panel of genetic markers are selected for each chromosome. The genetic markers (usually

microsatellites) are chosen from a *genetic map* which defines the genetic distance (in cM) between markers. The normal spacing between markers is around 20 cM resulting in a total of roughly 400 markers to cover the whole genome in a first pass genome scan. Each marker loci is genotyped in each individual. Alleles are then compared between parents and offspring to determine, if possible, which alleles were transmitted from each parent. The pedigree is then analyzed to see if particular alleles tend to be co-inherited with disease. In the case of complete linkage, all the offspring that have the disease also inherited the same marker allele from the affected parent, whilst all the non-affected offspring inherited the alternative marker allele from the affected parent. In such case all inherited chromosomes are referred to as *non-recombinant* and suggest that marker locus is located very close to the disease gene. In the case of no linkage, the co-inheritance of marker alleles and phenotype is random. Since there are two alleles at a marker locus and only one is passed on to the offspring, the chance is 50% that a specific allele will follow along with the disease phenotype. Thus when there is no linkage between the disease phenotype and the marker locus, there should be an equal number of *recombinant* and *non-recombinant* haplotypes in the offspring of parents who are heterozygous for the disease phenotype (assuming an autosomal dominant model of disease). There are statistical measures for linkage that quantify the likelihood that the marker is located close the disease gene, and thereby reveal the approximate position of the gene relative to the map position of the marker.

Recombinant:

In linkage analysis, a recombinant haplotype is scored when the disease phenotype and a particular marker allele do not co-segregate.

Ideal conditions for linkage studies would include access to large families (many generations of many offspring and thus many meiosis to examine) where the underlying genetic cause of the disease is descendent from the same mutational event (single founder effect). Also the presence of the disease gene should always give rise to the disease (highly penetrant) resulting in a tight genotype-phenotype relationship. Linkage studies have been extremely successful in the positional cloning approaches, especially in the case of simple Mendelian or monogenic diseases.

Association studies

For many complex diseases, families can be difficult to find or recruit, so an alternative approach is to perform population-based association analysis (figure 6). The traditional design of which is called a *case-control study.* In this approach a population is chosen and two subsets of samples are collected; A group of people having the disease (the

cases), and a group of people that do not have the disease (the controls). To maximize the chances for success, it is important to "match" the two groups as much as possible in regard to age, sex, ethnic background, and any environmental factors so that the only known difference between the two groups is the diagnosis. In this way the study is molded to focus on the unknown (hopefully genetic) differences between the two groups.

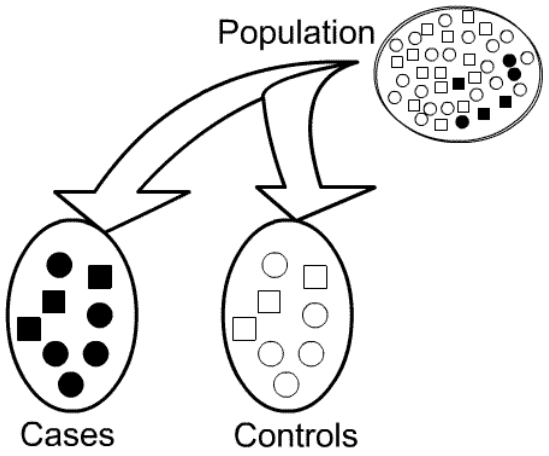


Figure 6: Design of a Case-Control association study.

The next step involves selection of genetic markers located in or around genes that are suspected to be involved in the disease. These *candidate genes* are most often implicated by one of two means. If biological evidence suggests that a certain protein or biological pathway is crucial for disease development, then a hypothesis could be that mutations in the gene that codes for the protein(s) are the underlying cause of the disease. Thus these genes are *biological candidates*. Alternatively genes can be candidates simply because of where they are located on a chromosome. These *positional candidates* are usually selected from chromosomal regions that have been implicated through positive findings in linkage studies.

Once the polymorphic marker is chosen, it is genotyped in both cases and controls. Allele and genotype frequencies are tallied and compared between the two groups. If a particular allele (or genotype) is much more abundant in cases compared to controls, then that allele is implicated in the disease. A statistical test, such as a chi-square (χ^2) test can be used to measure the strength of the correlation. A strong correlation, however, does not necessarily mean that the tested allele directly influences the disease phenotype. One alternative explanation would be that the neutral marker allele happens to be on the same haplotype as the pathogenic variant. In such case, it is the *linkage disequlibrium* (LD) between the marker allele and the pathogenic allele that accounts for the strong association.

Association study:

In simple terms, to test if a DNA variant is more common in a group of affected individuals compared to non-affected individuals.

A number of errors can also explain association signals, several of which stem from improper selection of cases and controls. A condition termed *population stratification* occurs when the patient group comes from a different ancestry than that of the controls. In such a scenario, allele and genotype frequency differences can be due to the individuals origins rather than disease status. Attempts to circumvent this problem are centered on shifting from a population based study to comparisions within various family structures [46]. A related potential problem is simply the random chance that samples chosen do not accurately represent the population at large. This combined with performing many association studies in the same patient and control material can lead to spurious correlations that are strictly due to chance. These *Type I errors* can be reduced by increasing the numbers of cases and controls studied and/or increasing the stringency of the statistical test.

Population stratification:

A source of error in association studies that is caused by the presence of multiple subgroups (with different allele frequencies) within the cases or controls.

Combined approach

Most of the successful attempts in mapping risk factors of complex disease have been made through a combined linkage and association approach. Initial mapping efforts by linkage often result in rather large chromosomal regions, in the order of 1 to 10 Mb in size, being implicated in the disease. Reduction of the putative region or *"fine mapping"* is often accomplished by follow-up association studies of the candidate genes in these regions. Isolation of risk alleles for both late-onset Alzheimer's disease (LOAD) [47] and Crohn's disease [48] have been cloned in this manner.

2) Which genetic markers to choose?

There are two major facets to this question. The first is the decision of the *type* of polymorphic markers to use and in contemporary research this boils down to the option between microsatellites and SNPs. Once this choice is made, the question can be redirected to inquire as to which of the microsatellite markers or which of the SNPs are most beneficial for genetic studies of disease.

The choice between SNPs or microsatellites is highly debated, and at the core of the matter are the comparative properties of each type of polymorphism. Microsatellites, for example, are extremely well suited for linkage studies where the

characteristically high heterozygosity makes the tracking of alleles through pedigrees much easier. The drawback with microsatellites comes mostly from complications involved in genotyping. Compared to SNP genotyping, there are remarkable few alternative methods for microsatellite genotyping. In all but the most advanced labs, there are a number of manual steps involved which in turn increase the risks for human error. In addition, interpretation of the experimental data can be problematic due to "stuttering" by the enzymes that are part of the genotyping chemistry, and the rather elaborate software required to call the genotypes [49]. Microsatellite genotyping is also comparatively time consuming and costly.

SNPs are typically only di-allelic and thus have a much lower heterozygosity and information content than most microsatellite markers. This binary character, however, has its benefits. A wide number of platforms have been adapted to take advantage of plus-minus character of calling SNP genotypes. Streamline implementations are comparatively quick and claim extremely high accuracy. Proponents of the use of SNPs are also quick to point out the ubiquitous distribution of SNPs in the genome, and emphasize the density of SNPs, in reference to the distance between markers, is much higher than for microsatellites. In addition, the mutation rate in SNPs ($1x10^{9}$) [50, 51] is roughly 10,000x lower than the estimated mutation rate of microsatellites ($1x10^{5}$) [52]. Thus, mutational events are less likely to interfere with association or LD studies when using SNPs. The lower mutation rate of SNP would also imply that higher frequency SNPs are very old loci and are more likely to be common to populations around the world compared to microsatellites.

If microsatellites are chosen, then marker selection is usually customized to the goals of the study. For instance, a first linkage study may require a relatively sparse collection of markers to cover a chromosomal region. If linkage is detected, additional microsatellite markers can be added to those regions of interest to help focus in on the linkage signal. Markers can be added and typed until no more recombination events are detected in the region. The number of markers needed is at least in part a function of the number of individuals (meioses) in the linkage study.

When it comes to SNPs in complex disease, it is important to remember that mutations that cause disease are created by the same mechanism used to create polymorphism. If the severity of the pathogenic change is low enough or has its effect at a late stage in life, then risk alleles can spread in a population and reach polymorphic levels. Selection of SNPs for association studies can be shaped to include those SNPs that are most likely to directly influence the disease phenotype [53]. For instance, SNPs

that occur in exons can either change the encoded proteins amino acid sequence (and are thus termed non-synonymous substitution) or not (synonymous substitution). Altering the protein sequence may well affect function, so there might be a case to use non-synonymous SNPs over synonymous for association studies. Similarly, SNPs that occur in the promoter regions might affect expression levels and so play a more active role in disease etiology than SNPs found in the middle of introns. A sort of pseudo-hierarchy can be constructed based on whether or not the individual SNP is likely to have a direct impact on the disease phenotype (figure 7).

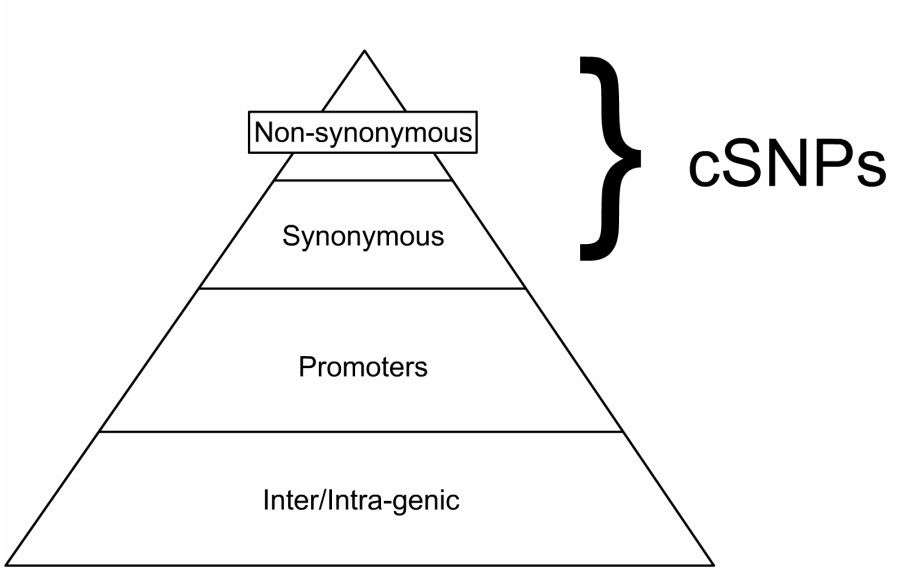


Figure 7: Hierarchy of SNPs based on the potential to directly affect disease phenotype.

The exact order of this hierarchy can be debated. Promoter variants may well have a greater influence on complex disease than synonymous or even non-synonymous coding SNPs (cSNPs), but the main point of this discussion is that not all SNPs are equal, and selection of SNPs may have a bearing on the success of the study.

One final consideration when it comes to SNPs is the question of allele frequency. It is tempting to use SNP markers with relatively high minor allele frequencies for several reasons. In discovery efforts, these SNPs are the first to be detected and require the fewest numbers of individuals to be screened. Also high frequency loci are less likely to be population specific, and thus the same marker can be used for studies in many populations. In addition there is a theory called the common disease common variant (CDCV) theory[54, 55] that postulated that many of the common diseases are caused by combinations of common variants, and thus much of the disease etiology can be explained through the interaction of high frequency alleles of SNPs.

Counter arguments are many and varied; with a common thrust being that complex disease may well be caused by numerous rare variants rather than combinations of common variants [56]. This is known to be the case in many monogenic diseases, a trademark example of which is pigmentary retinopathy where more than 600 disease causing mutations at 55 loci have been identified (www.sph.uth.tmc.edu/RetNet/disease.html). If complex disease, where selection pressure is presumed to be low to non-existent, exhibits a lot of allele or locus heterogeneity, then standard association studies using any SNP markers is not likely to work [57]. The true allelic spectrum of human disease [58], especially complex disease, remains a mystery for the time being and will no doubt be a source of controversy for years to come [59].

3) *How many SNPs do I need to assure I find the genes involved in my disease?*

As eluded to in the previous section, the absolute number of SNP markers that need to be tested in order to find the genetic determinants of complex disease is difficult to answer. Some daring individuals [60, 61] have proposed that it might be feasible to perform *linkage disquilibrium mapping* over the entire human genome in one experiment. The design of this very large association study would involve tens [62] to hundreds of thousands [63-66]of SNP markers to be genotyped in a large set of cases and controls. No candidate genes are selected *a priori*, so like linkage analysis this is a reverse genetics approach and consequently relies on phenomena like LD to indicate the vicinity of disease genes in the genome.

To minimize the number of SNPs needed in such an approach, large efforts are underway to determine the amount and patterns of LD in the human genome through both modeling [67] and experimentation [30, 64, 65, 68, 69]. The idea is to map regions of the genome where LD is strong, and then classify them into various *hapotype blocks*. In such regions, the total allelic diversity in the block is represented by a handful of major haplotypes. By carefully selecting a few key polymorphic markers, the major haplotypes can be captured with a minimum amount of genotyping effort. The markers that define major haplotypes in a region are usually referred to as haplotype tagging markers (htSNPs) [70]. This haplotype information could also have a bearing on association studies in candidate genes, as a set of htSNPs could be used to quickly screen genes thought to be

Haplotype block:

Stretch of DNA that shows high linkage disequilibrium and low haplotype diversity

involved in disease. A cautionary note, however, is that these theories are fairly recent and have yet to be confirmed in actual disease studies. There are currently several different ways to measure LD [37, 71, 72] and no consensus for how to define haplotype blocks or choose htSNPs. Evidence thus far in the construction of the human "HapMap" [68, 73] is that there seems to be some consistency in LD patterns among global populations [30, 68, 74, 75]. It seems likely that a HapMap would prove interesting for population genetics point of view, but what remains to be seen is if this information can be used for understanding the relationship between genetic variation and disease [45].

4) Which population to study?

One important consideration in mapping disease genes is defining the population of study. Conventional use of geographical boundaries such as national borders depicted on a map may not be optimal to ensure that the study population is genetically homogeneous. Other factors such as ethnic origin, language, or possibly even religious orientation can help refine populations in terms of shared history, and thus increase the chances that the individuals share the same genetic background. Partitioning along these additional criteria have led to the definition of "population isolates" such as that of the Saami in northern Scandinavia [76], the Finnish population[77], or the Mennonite populations [78]. Besides reduced genetic complexity, other potential advantages of population isolates could be the standardization of shared environmental factors such as diet or exercise and/or exposure to infectous disease [79]. Some isolates such as the Icelandic population have extensive genealogical data [80] which can further assist in the reconstruction of pedigrees and tracking founder effects for disease.

The debate persists on whether population isolates will be as fruitful in the identification of risk factors for complex disease as it was for finding mutations in rare disease [81, 82]. If, indeed, there exists extended amounts of LD in these populations, a two-punch approach may be effective for mapping disease genes. A first step might be to perform the studies in the isolated populations for initial mapping followed by replication in more outbreed populations for fine-mapping [83]. As preliminary data is collected, however, there appears to be only a modest elevation of linkage disequilibrium in isolated populations such as Finland or Sardinia compared to neighboring outbreed populations [84, 85]. In such case isolated populations may not deliver the expected advantages of increased linkage disequilibrium. Other concerns with isolated populations is simply being able to collect sufficient numbers of

individuals to make the study statistically meaningful, and that in the case of genetic heterogeneity in disease susceptibility that results obtained in isolated populations may not be pertinent to the cause of disease in other populations.

5) How many individuals?

A very general statement would be "the more, the better". Both linkage studies and association studies benefit statistically from the inclusion of large numbers of individuals. Extended pedigrees can assist in narrowing linkage signals to smaller chromosomal intervals in linkage studies. In association studies, large samples reduce stochastic errors that can be introduced into the study through random sampling. The scale or number of individuals studied also has a direct relationship to statistical measure of *power*. In principle, power calculations are a statistical attempt to relate sample size with the magnitude of the affect that can be detected. Values range between 0 to 1 (or 0-100%), and the larger the number the higher the probability of finding an association should it truly exists. A typical study design might aim for 80-95% power to detect a statistically significant association of a variant genotype that results in a doubling of the risk for disease [86]. Ideally power would be used prior to performing the study to determine the number of samples that need to be collected. However, quite frequently in genetic epidemiology the power formulas are used to estimate the maximum genetic contribution that can be detected given the (normally small) number of samples analyzed. Minimum sample sizes for detecting a doubling of risk with 80% power might be in the range of several hundred cases and controls [87, 88]. There are a number of fairly straight forward methods for calculating the power of a particular study [89], and there are several sources of helpful tools available online (http://members.aol.com/johnp71/javastat.html#Power).

Power:
The probability of finding a correlation if it truly exists.

Aside from the absolute total number of individuals in the study, a second consideration is the distribution of cases and controls. Association studies rarely have the exact same number of cases and controls, and most often there are more controls. This is largely to that cases tend to be more difficult and costly to collect. If the cases are to be as homogenous as possible, it is important to establish diagnostic criteria that are strictly followed for recruitment of patients. If studies from different groups are to be comparable, the same guidelines for diagnosis need to be used for all studies. One potential way to increase the chances of finding genetic risk factors in complex disease

is either to select only cases that show a severe phenotype, or perhaps select a sub-phenotype for analysis in association [90]. The thought is that more discrete phenotypes may be caused by genetic risk factors that are more tractable than in the disease as a whole.

6) Are my results significant?

"An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts – for support rather than illumination"

Andrew Lang

It is important to differentiate between statistically significant results, and results that are biologically significant. This rather abstract thought is based on the fundamental difference between statistical and biological evidence. Most statistical tests used in disease studies are tools that measure how likely we are to obtain the observed result as compared to chance. It is up to the scientist to determine whether or not the correlation is valid. To illustrate the point, a study may find a statistically significant correlation between taking a pill and reduction of drowsiness. If the pill was a placebo, then the reduction in drowsiness wasn't a direct result of the pill itself but rather the individuals' perception of what the pill should be doing. Furthermore, even if the statistics do correctly identify a cause-effect relationship, it is still up to the scientist to determine if the result is biologically meaningful. As an example there may be a statistically significant correlation between taking a certain drug and decreasing blood pressure by 1 unit. A statistical measure may indicate that there is a 1:1 billion probability that this was a result of chance, whereas a doctor may well say that this result is not interesting biologically.

Bearing this in mind, it may be easier to evaluate the role of statistical methods used in examining the numbers generated by linkage and association studies. Statistical tests for linkage are usually based on likelihood, and through a series of variables compare the observed *recombination fraction* (θ = the ratio of recombinant to non-recombinant haplotypes) to the expected recombination fraction assuming no linkage ($\theta = 0.50$). When performing a genescan using a model-based or *parametric* linkage analysis [91], the result is usually presented as logarithm of the odds score (*LOD score*), denoted Z in the literature. If Z

Parametric LOD-score method:

A form of linkage analysis that requires specification of disease model parameters such as penetrance, mode of inheritance, and disease allele frequency.

≥3.3, it is typically accepted as strong evidence for linkage, while Z ≥1.9 is sufficient for suggestive linkage [92]. Non-parametric studies have a slightly higher threshold for significance, with a Z ≥ 3.6 for strong linkage, and a Z ≥ 2.2 for suggestive linkage.

A common statistical measure of association in case-control studies is the calculation of the *odds ratio (OR)*. Basically the OR tries to calculate the *magnitude* of differences in allele or genotype frequency between cases and controls. To assist in OR calculations, a good way to organize data is to create a two-by-two table like the one given below (figure 8). Exposure, as presented in the diagram, is a universal epidemiological term that can be substituted for allele or genotype counts.

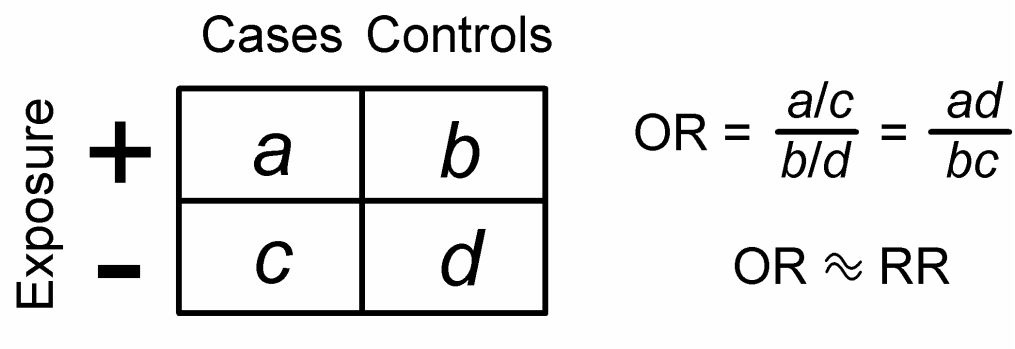


Figure 8: Two-by-two contingency table. The generic formula is given for how to calculate the odds ratio (OR) which in case-control studies is an estimation of the relative risk (RR).

The OR calculation as described above gives the *odds of having the disease* given the exposure, compared to not having the disease and the exposure. Thus it is an approximation of the *relative risk (RR)* or the *risk of acquiring the disease* dependent on the exposure. In case-control studies, RR cannot be calculated directly, but with sufficient numbers of cases and controls, the OR is a fairly good approximation of the RR. A relative risk of 1 means there is no difference in disease risk regardless of exposure, more than 1 indicates exposure increases the risk of disease, while less than 1 indicates a decrease in risk such as is the case with protective factors.

Beyond calculating the statistical magnitude of the association, it is also necessary to calculate the *validity* or accuracy of the measurement. The most common measure given in the literature for this is the probability value (p-value). By definition, the p-value is the probability of obtaining as extreme or more extreme result (value of association) based on chance. P-values are given as a fraction, and the standard acceptance level for significance is a p≤0.05, indication the chances

p-value:

The probability that the observed result happened by chance.

were 1:20 or less of obtaining the result. One statistical route to calculating the p-value is to perform a chi-square (χ^2) test. After calculating the χ^2 value, conversion to a p-value can be done through simple use of a reference table. A higher χ^2 value will result in a lower p-value.

The p-value threshold of 0.05 is the generally accepted for the analysis of one single test (i.e., testing genotypes of 1 SNP). If multiple tests are performed, by examining multiple SNPs for example, then it is more likely to arrive at a significant result by chance. Accordingly, the p-value must be made more stringent to account for the increased number of tests. The Bonferonni correction stipulates that for each additional independent test, the threshold for significance should be raised and done so by dividing the significance value by the number of tests performed. Hence, testing two SNPs in an association study would require a p-value of (0.05) / 2 or 0.025 to be considered statistically significant. Many consider the Bonferroni correction for multiple testing to be too conservative [93,94], and other methods of correcting for multiple testing based on permutation have been proposed [95].

In summary, there are a number of non-trivial considerations that need to be addressed in order to raise the chances for success in mapping disease genes. This is far from an exhaustive list, and the number of issues will continue to grow as the search for tractable genetic elements in disease expands to include more and more difficult things to find. The most difficult decision, in my opinion, is deciding on which diseases to study, and which ones to leave alone. Several theoreticians speculate that for some complex diseases, the reason why we do not find genetic risk factors is not in our low-throughput technology but in our biology [45]. That in essence, increasing the number of drawers we look in will not increase our chances of finding our glasses, if our glasses are on our head. Sorting out which phenotypes are worth studying is much easier in retrospect, but still a critical decision to make prior to the start of a genetic mapping effort.

Case study - Alzheimer's disease

One complex disorder in man where there has been varying levels of success in finding genetic risk factors is Alzheimer's disease (AD). As alluded to earlier, the incidence of AD has been increasing steadily and as of the few decades, AD has become leading the cause of dementia and the fourth leading cause of death in the elderly of western societies [96]. This has spurred massive efforts in the research community to attempt to find risk factors involved in AD [97].

Clinical symptoms and diagnosis

As first described by the Bavarian psychiatrist Alois Alzheimer in 1906 [98], the cardinal symptoms include progressive memory impairment, disordered cognitive function, and altered behavior including paranoia, delusions, and deterioration of language skills. Clinical diagnosis of the disease has since been refined in efforts to distinguish AD from other types of senile dementia. AD is now considered a progressive disease with several intermediate stages of severity. The earliest stage is almost indistinguishable from normal aged forgetfulness [99] with symptoms such as difficulties in concentration or remembering where things have been placed. As the disease progresses, cognitive abilities continue deteriorate and complex activities such as balancing a checkbook become troublesome but the individual can still perform most tasks in daily life. Beyond this mild AD stage [100], cognitive decline advances and motoric and/or sensory functions can become affected. Coordination can be impaired to the point where the individual has trouble walking and eventually even talking. People suffering from severe AD become very susceptible to infectious diseases such as pneumonia which can lead to death, however some patients show no other cause of death than AD [101].

Although not entirely exclusive to Alzheimer's disease, the presence of two types of brain lesions are commonly found in the AD patients. These neuritic plaques and neurofibrillary tangles commonly occur in the limbic and associated cortices [102]. Neuritic plaques are spherical lesions that contain extracellular deposits of amyloid-β (Aβ) protein. A form of the protein ($A\beta_{42}$) is particularly prone to aggregation and is often the major constituent of the plaque core [103]. Neurofibrillary tangles are most often composed of pairs of filaments wound into helices. These paired helical filaments (PHF) are composed of the microtubule-associated protein called tau [104] that become insoluble and precipitate, presumeably through hyperphosphorylation [105].

Genetics

As mentioned previously, there seems to be a significant genetic component in AD etiology [44]. In addition, there is a severe sub-type of the disease called early-onset Alzheimer's disease (EOAD) or familial Alzheimer's disease (FAD) that appears to be Mendelian in nature. This is a fairly small fraction, however, with 95% of all AD cases having an age of onset of ≥ 65 years of age and showing characteristic complex inheritance patterns. The search for genetic factors in this late-onset Alzheimer's disease (LOAD) compared to FAD has proven much more difficult.

Several genetic risk factors for FAD have been discovered thus far through genetic studies. A first genetic clue to the chromosomal location of a AD disease gene came with observation that patients with Down's Syndrome (trisomy 21) also develop some of the classical signs of AD [106]. When amyloid precursor protein (APP), the precursor of the $A\beta_{42}$ fragment found in neuritic plaques, was cloned to the same region on chromosome 21 that was implicated in several linkage studies, it became an obvious candidate. However, APP mutations are rare and mutations in this gene have only been found in some 25 FAD families worldwide [107]. Further linkage studies pointing to Chromosome 14 [108] led to the discovery of Presenilin I (PSEN I) [109]. Some 75 mutations in this gene have been identified to date making PSEN I the most important genetic risk factor in FAD identified so far. A homologous protein, denoted Presenilin 2 (PSEN 2), has also been implicated in FAD, and to date only 3 mutations in this have been found to cause FAD.

Two lines of evidence eventually tied Apolipoprotein E (APOE) to LOAD. Studies of the cerebral spinal fluid of AD patients found that Aβ peptide recovered from this source was often aggregated together with the APOE protein [110]. The mapping of APOE to chromosome 19 fit nicely with AD linkage studies results that implicated a disease susceptibility locus in the same region [111, 112]. A pair of non-synonymous SNPs [113] was quickly tested and of the three haplotypes observed between these two SNPs, the haplotype or allele denoted "APOE-ε4" was found to be overrepresented in LOAD cases compared to controls [47]. The APOE-ε4 association has been replicated in numerous populations around the planet, and is one of the very few success stories for the candidate-gene, case-control studies in complex disease.

Current treatments

At the time of writing this thesis, a cure for Alzheimer's disease is still out of reach for modern medicine. On the hopeful side, however, several medicinal therapies have been

shown to slow down the progression of the disease. The pharmaceutical agents are all cholinesterase inhibitors sold under the trademarks Aricept (Esai), Exelon (Novaritis) and Reminyl (Janssen). These drugs are designed to increase cholinergic activity in the brain and have proven to modestly improve cognitive function [114]. A more promising drug, currently approved only in Europe, is a substance called memantin under the trademark Ebixa (Lundbeck). This drug works to restore the normal function of the NMDA-receptor which can be constitutively activated as a result of AD. One last treatment that has been researched is a potential vaccine against Alzheimer's disease. The mechanism involves stimulation of the bodies own immune system to react against Aβ, the main component of the AD related neuritic plaques. In animal studies, vaccination has lead to a stoppage and in some case reversal of Aβ deposition in the brain [115,116] however all drug trials in humans have been stopped because of severe side-effects (development of meningitis and/or encephalitis).

SNP TECHNOLOGIES

Since the time I entered the field in 1997, research in molecular genetics technologies has experienced a veritable golden age of innovation. In parallel, SNPs have become exceedingly popular as markers in disease studies. The direct consequence of these trends has been the invention of an extensive battery of strategies to distinguish SNP alleles. Although these screening techniques differ widely in design and construction, they all share the same goal of providing the research community with viable alternatives for extracting genotype information from patient DNA.

In order to keep pace with the ever-widening variety of techniques, a number of high quality reviews have been produced concerning methodologies for SNP screening [117-126]. A common approach taken to assist in understanding the differences and similarities of each is to break the techniques down into three fundamental concepts. Namely; the ***reaction principle*** by which the alleles are distinguished, the ***reaction format*** defining milieu in which the reaction takes place, and the ***detection mechanism*** through which the allele specific products are visualized. In the following sections I will briefly summarize these three modalities and then show how they are combined in a number of predominant methods currently in use in the field.

Discovery vs. scoring

Before diving into classification schemes, a clear distinction can be made between methods designed to discover *new* SNPs, and those designed to score *known* SNPs. To date, no method has provided an ideal solution for both tasks. Classic discovery techniques, such as single-stranded conformation polymorphism (SSCP) [127] and denaturing gradient gel electrophoresis (DGGE) [128], pre-date much of the DNA sequence data we have today and consequently require minimal sequence information prior to performing the assays. Both SSCP and DGGE can be used to screen DNA fragments for dissimilarity, however to identify the actual sequence differences requires additional methodologies to be employed. Therefore, the vast majority of all new SNPs are characterized through a penultimate step of DNA sequencing.

Reaction principles

The first step in converting a DNA sample into a useful format for genotyping is performing polymerase chain reaction *(PCR)* [129] on the sample. Regardless of which allele discrimination reaction is implemented, all current methods work best if PCR is performed first. The effect of running a PCR on a sample is a high enrichment of just the specific fragment of DNA that is to be interrogated (improving specificity) as well as providing many more DNA molecules for study (improving sensitivity).

PCR:

Polymerase chain reaction: a method for making millions of copies of a selected fragment of DNA

Once the PCR product (that spans the SNP loci) is created the next step is to find out which alleles are present in the sample. The four common ways are through hybridization, digestion, extension, or ligation (figure 9).

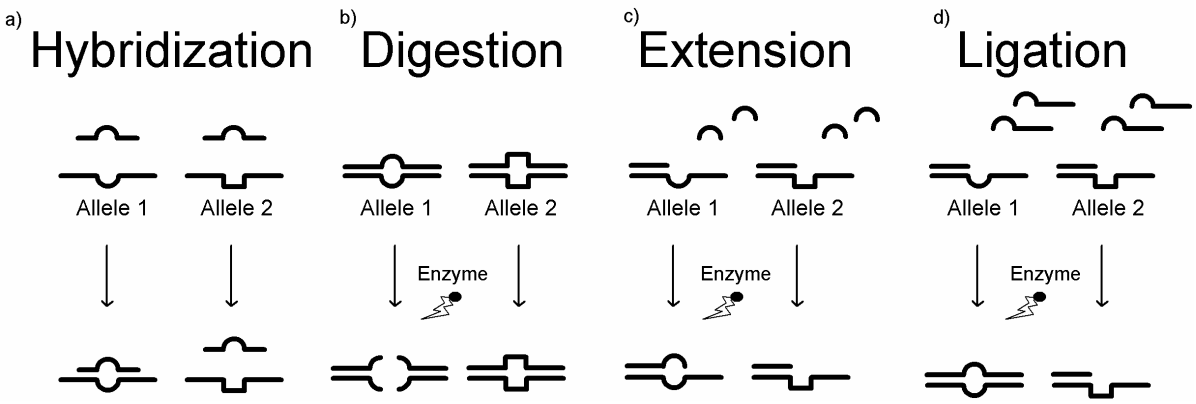


Figure 9: **Reaction principles for allele discrimination. For hybridization, alleles are scored through differences in how the target alleles interact with a probe. Digestion uses enzymes that only cut the DNA if the appropriate sequence or structure is present. Extension uses polymerase enzymes to add single nucleotides only if the appropriate sequence is present in the target. Ligation uses DNA repair enzymes to link two probes together only if the appropriate sequence is present in the target.**

Hybridization

Hybridization (figure 9a) takes advantage of the double-stranded nature of DNA. Given the sequences are *complementary*, two single stranded DNA molecules placed together will stick together in a structure called a DNA duplex. The stability of the DNA duplex is dependant on a number of factors such

Base complimentarity:

DNA is composed of anti-parallel strands of nucleic acids. The sequences are" complementary" as an "A" base in one strand matches specifically with a "T" in the other, and similarly a "C" base matches a "G" base.

as DNA sequence, concentration, length and secondary structure [130] as well as reaction conditions such as ionic strength, pH and temperature [131].

To differentiate between alleles, hybridization assays involve the interaction of single-stranded DNA targets and an allele-specific oligonucleotide probe. Targets that are 100% complementary to the probe can form a more stable DNA duplex than targets that contains a non-complimentary or "mismatched" base. In other words, the target DNAs that contain the "match" allele will form more stable duplexes with the probe than will "mismatch" targets. Scoring SNPs thus becomes a matter of adjusting the stringency of the reaction conditions and monitoring for the presence or absence of a DNA duplex.

It should be noted here that hybridization is the only one of the four allele discrimination tactics that does not require enzymes beyond the initial PCR amplification step.

Digestion

The digestion process (figure 9b) relies on specialized enzymes that accurately recognize specific DNA sequences. These enzymes act as biological scissors, cutting double-stranded DNA whenever a particular series of bases is present in the target DNA sequence. Many *restriction endonucleases* have been identified for many different (usually palindromic) DNA sequences. Allele discrimination is possible if the SNP locus occurs in the recognition sequence for one of these enzymes. One allele should give the correct recognition sequence and allow for digestion, while presence of the other allele should disrupt the recognition sequence and prevent DNA cleavage.

Extension

The extension principle depends on enzymes that add bases to DNA sequences. Using a single-stranded target, these *polymerase* enzymes can be used in two similar ways for allele discrimination. The first way is to hybridize an oligonucleotide to the target so that the 3' end of the oligo falls on the polymorphic position in the target. Since the polymerase requires the 3' end of the oligo to be complementary to the target, alternative alleles can either allow or prevent the polymerase from adding nucleotide bases. In this scenario, SNPs are scored by monitoring *if* DNA extension has occurred.

Alternatively, an oligonucleotide can be hybridized so that the 3' end of the oligo is adjacent to the polymorphic position. Modified nucleotide bases can be added to the extension reaction that both allow for unique detection of each type of nucleotide (i.e.,

A, C, T, or G), and also prevent to polymerase from adding more than one base. For these single-base extension (SBE) techniques, allele discrimination occurs by monitoring *which type* of base is added to hybridized oligonucleotide (figure 9c).

Ligation

The ligation-based assays involve enzymes that join DNA fragments together (figure 9d). For these *ligase* enzymes to link two oligonucleotides, the probe pair must be annealed adjacently on the target DNA. In addition, the ligase requires that the bases at the probe-probe junction be perfectly complementary to the target. If the SNP position is at the junction, then one allele will fulfill this requirement and the oligo probes can be ligated, while presence of alternative allele will result in non-complementarity at the junction and thus prevent ligation.

Reaction formats

All of the reactions mentioned above take place in aqueous solutions. Upon completion of the allele-discrimination reaction, the next step is to prepare the allele-specific products for detection.

Homogeneous reactions

Genotyping platforms that remain in solution throughout the entire genotyping procedure are termed *homogeneous* assays. The most streamline of homogeneous assays involve PCR, allele-discrimination, and detection in the same vessel and require no intervention beyond initial set-up. Other homogeneous methods may call for additional reagents to be added to the reaction vessel, but no separation or purification of products is required for product analysis.

Solid-phase reactions

In order to facilitate purification or separation of allele-specific products, many genotyping techniques are assisted by immobilizing the products to some type of surface. The support material is usually some type of plastic, polymer or glass, and implementations of these materials have resulted in a wide variety of support structures or forms. One simple support is merely the wall of a reaction vessel. The wall of a microtiter well, for example, can be modified so as to allow attachment of DNA. Anther type of support would be planar structures such as glass slides, silicon chips, or

even flexible membranes which are used to arrange samples onto a flat surface. These *DNA arrays* can be simple or complex, with the density of features reaching up into the tens of thousands per square centimeter. Another type of support is latex beads. These small spherical bodies can be modified to attach specific DNA molecules. DNA-coated microparticles or "microspheres" have advantages in that DNA molecules are collected onto discrete surfaces, however the particles are often neutrally buoyant and can still freely interact with the reagents in solution. Although not technically solid-phase immobilization, agarose or polyacrylimide gels are commonly used to separate allele-specific products from reactants. These materials provide a matrix for separation of DNA molecules based on charge and/or molecular weight. Homogeneous and solid-phase alternatives are depicted in figure 10.

DNA array:

A 2-dimensional arrangement of DNA samples on a solid-surface.

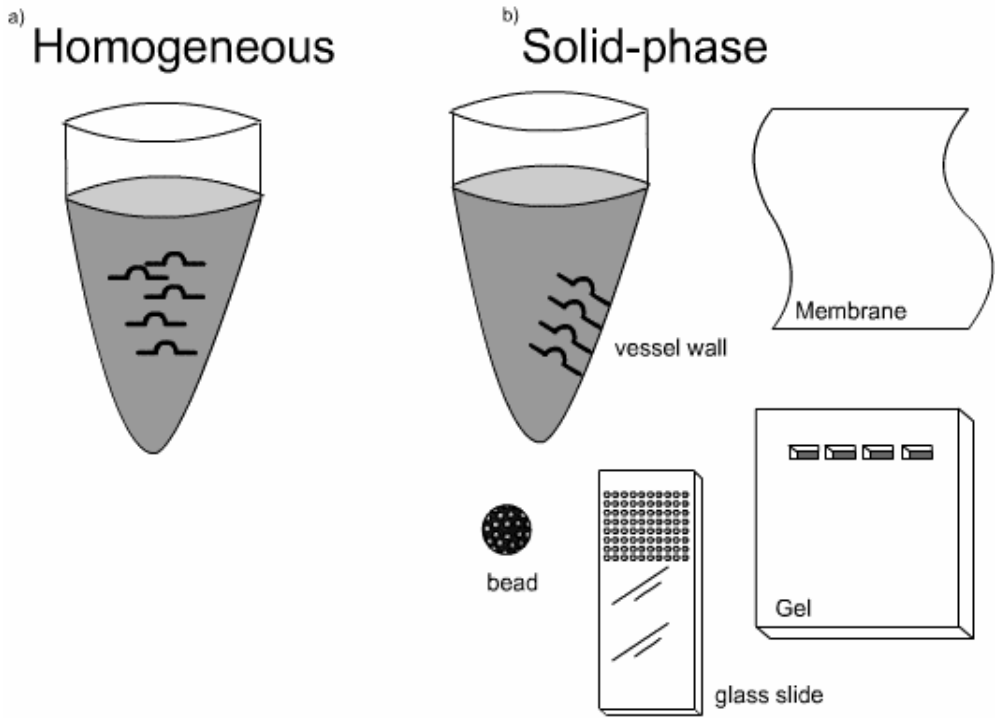


Figure 10: Homogeneous assays take place in solution. Alternatively, sample can be processed on a wide variety of solid or semi-solid surfaces.

Detection mechanisms

DNA is a wondrously small molecule. The width of a double-helix is about 2 nm, or about 1 million times smaller than the width of a pinhead. As a result, direct visualization of DNA molecules is extremely difficult even with the most sensitive techniques such as scanning transmission electron microscopy (STEM) or atomic force microscopy (AFM). It is therefore necessary to modify DNA in some way in order to track where it is, or what the sequence is. Early research employed radioactive particles for this purpose. Different radioactive isotopes such as ^{3}H, ^{32}P or ^{35}S can be built into a growing DNA chain and detected using x-ray film exposure (autoradiography) or through other radiation measuring devices such as a scintillation counter or a phosphoimager.

Although very sensitive, radioactive labeling of DNA molecules has been phased out as the detection method of choice. Modern genetic research is investing heavily into less toxic techniques of DNA detection. The results of genotyping procedures are now interpreted through examining changes in the mass, electrical conductivity, or the light-emitting properties of labels affixed to the allele-specific products.

Mass detection

Mass spectrometry (MS) can be used to measure small differences in molecular weight between allele-specific products. Specifically a technique termed matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectroscopy has been successfully employed in several SNP genotyping strategies [132-137]. The principle involves co-depositing a small amount of sample and matrix onto the surface of a metal target plate. A short laser pulse serves to desorb the DNA sample/matrix into the gas phase, and the ionized product is accelerated towards a detector. The mass of the products can be inferred from the relative time it takes for the product to travel from the site of ionization to the detector. The resolution of modern mass spectrometers can easily distinguish single base differences in short DNA sequences [138]; or mass-tags that differ by only a few Daltons (Da) [139, 140].

Electrical detection

One intriguing detection approach is through monitoring changes in electrical properties that occur as a result of DNA hybridization. A recent innovation enrolls allele-specific oligonucleotide probes that have been functionalized with gold

nanoparticles. Successful capture of the probes to target DNA molecules immobilized between two microelectrodes results in completing the circuit and a subsequent reduction in electrical resistance [141]. In parallel with the emerging field of nanotechnology, microelectronic approaches appear promising for detection of SNP alleles [142].

Detection of luminescence

Detection of emitted light, especially within the visible spectrum, has been adopted as a more versatile and benign alternative to the detection of radioactive isotopes. Many labeling systems have been developed to stain or tag DNA molecules in such a way that the light emission properties change as a response to the allele-discrimination event. Luminescence applications to genotyping include fluorescence, fluorescence resonance energy transfer (FRET), chemiluminescence and fluorescence polarization (FP).

Luminescence:
The emission of light from a substance that occurs from electronically excited states.

Fluorescence

In general terms, fluorescence occurs when a *fluorophore* absorbs light at one wavelength and emits light at a lower energy wavelength. At the quantum level, electrons are bumped up to an excited state and in the return to ground state excess energy is dissipated as fluorescence, or in some cases phosphorescence depending on the path the electron takes. The excitation spectrum is thus all the wavelengths of light that can excite the electrons of the fluorophore. The emission spectrum is all the wavelengths of light that the fluorophore can give off as the electrons return to groundstate. A common way to depict these excitation and emission spectra is illustrated in figure 11.

Fluorophore:
A molecule capable of fluorescence.

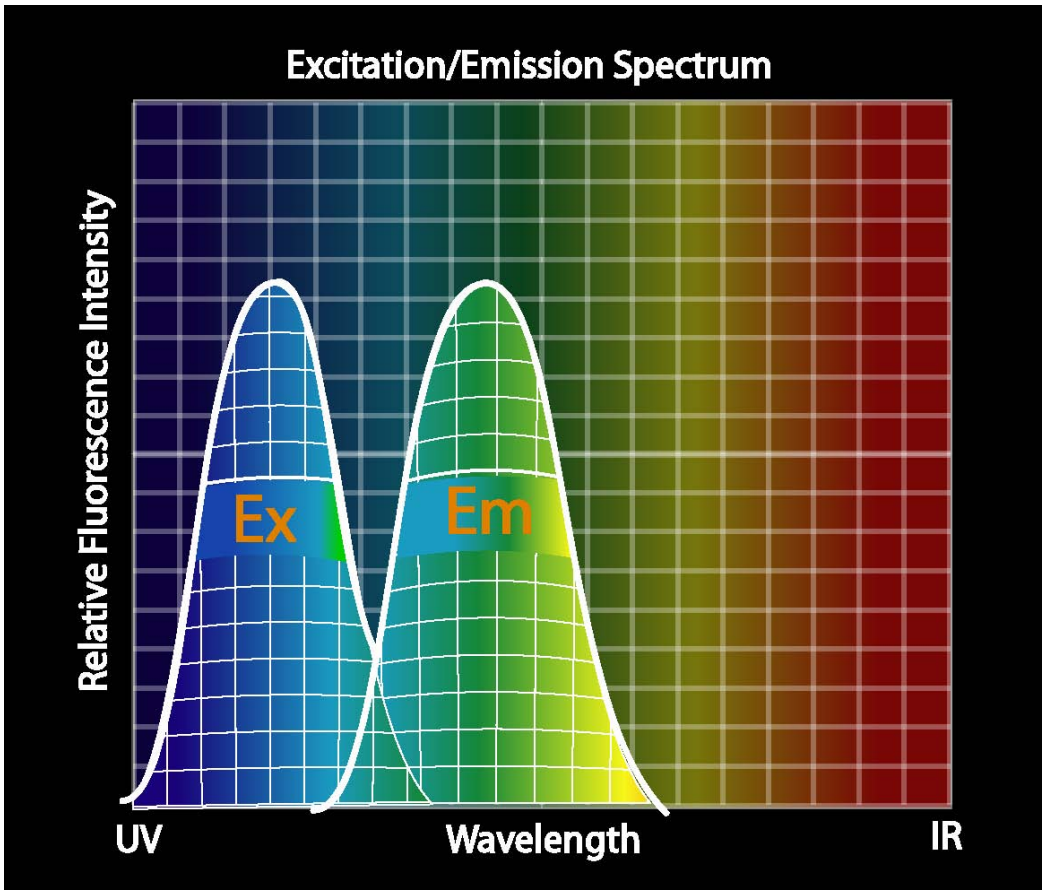


Figure 11: Excitation and emission spectrum. The X -avis represents the wavelength of light that is either absorbed or emitted, and the Y-axis is a relative measure of the fluorescence intensity.

Fluorescent compounds are abundant, a common place example of which is the yellow dye used to color Basset's wine gums. When exposed to UV light, the yellow wine gums appear to "glow". In genetic research, a number of fluorophores have been developed with any number of desirable characteristics. For example, intercalating dyes are fluorophores that bind specifically to double-stranded DNA. Sybr Green I ® (Molecular Probes, The Netherlands) is an exceptional example of this class that shows a 10,000x enhancement of fluorescence when bound to double-stranded DNA. Other fluorophores, such as cyanine or rhodamine dyes, can be directly linked to DNA molecules and thus allow for indirect tracking of the DNA molecules. Some of these, such as the Bodipy® series of dyes (also Molecular Probes, The Netherlands), have very sharp emission spectra. The advantage is that the individual emission of each dye can be distinguished using optical filters even from complex mixtures of the dyes. This allows for some genotyping strategies to examine more than one SNP at a time (so-

called multiplex analysis) by simply labeling each allele-specific reaction with a different color dye.

When selecting fluorescent labels, there are three additional properties that are good to consider. One is the *Stokes' shift*, or the distance between the absorption maximum and the emission maximum. A large Stokes shift is often preferable as the light emitted from the excitation source is easier to filter away from the fluorescent emission of the fluorophore. A second factor is the *extinction coefficient (ε)*. This variable describes the photon capture efficiency of the fluorescent dye, and values usually range from 10,000 – 250,000 $cm^{-1}M^{-1}$. Again, higher numbers are preferred as this indicates a better ability of the fluorophore to accept excitatory photons. The third property is the *quantum yield*. This describes the relationship between how many photons are absorbed compared to how many photons are emitted. This can be thought of as the fluorophore's efficiency in converting between incoming and outgoing photons with the maximum efficiency being equivalent to quantum yield of 1.

Fluorescence resonance energy transfer (FRET)

FRET is a phenomenon that occurs between distinct pairs of fluorophores. Given the correct conditions, a *donor* fluorophore can pass on its excitation energy to an *acceptor* fluorophore. Instead of emitting light, the donor resonantly transfers the energy to an acceptor that is close in proximity. The acceptor can then release a photon of light that corresponds to its emission spectrum or, in the case of dark "quenchers", dissipate the energy as heat. For FRET to occur, the emission spectrum of the donor must overlap with the excitation spectrum of the acceptor (figure 12).

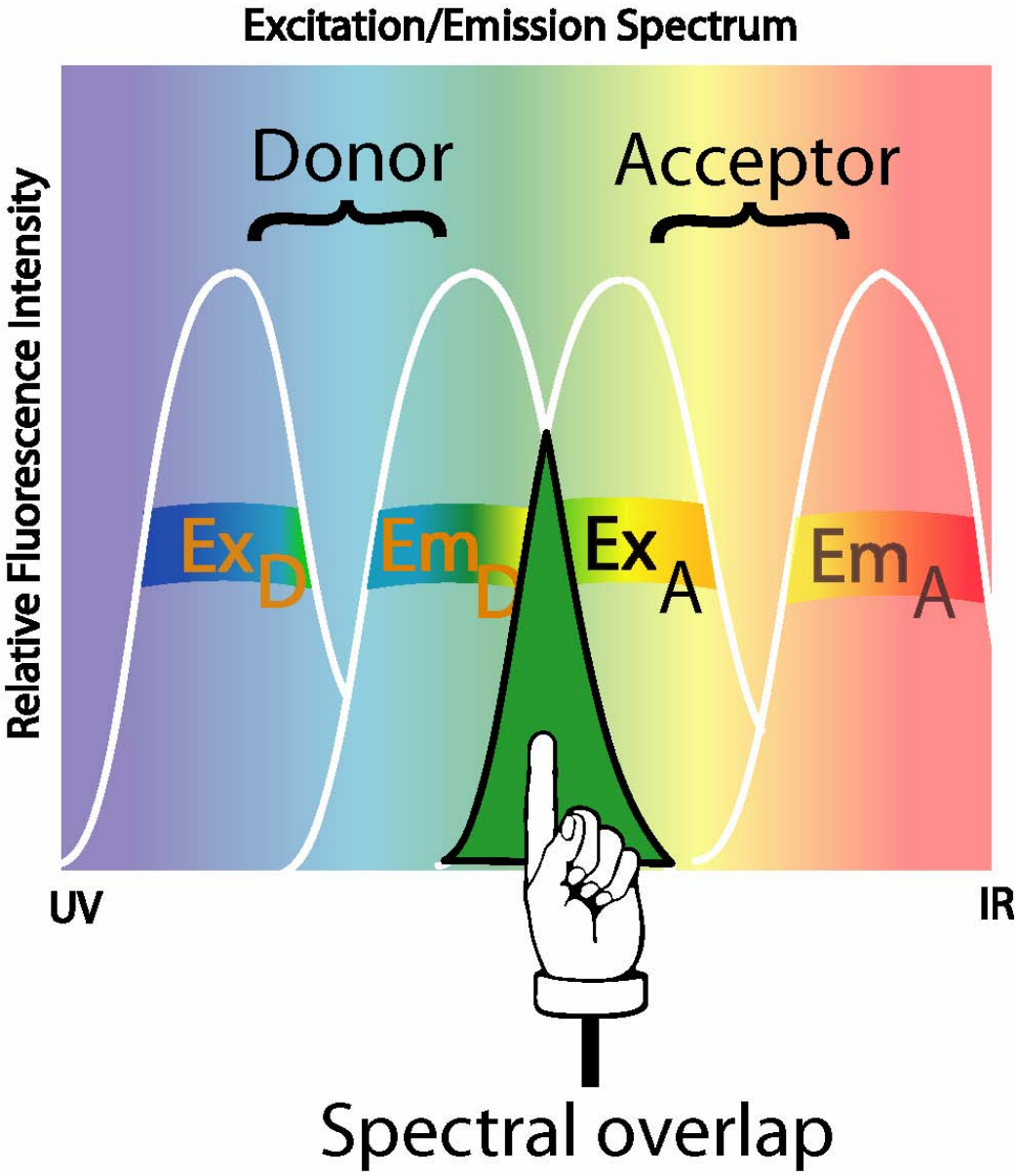


Figure 12: Excitation and emission spectra for two dyes that are spectrally compatible with a FRET reaction.

The efficiency of the energy transfer depends greatly on the spatial relationship between the donor and acceptor molecules. In general, the closer the acceptor & donor are together, the more efficient the energy transfer. As it so happens, the optimal distance for FRET to occur (called the Förster radius ~10-100Å) coincides well with the relative distances involved in DNA duplex formation. Therefore a number of FRET strategies in genotyping techniques are designed around the hybridization of DNA

molecules. A number of different spatial relationships of the donor and receptor are given in figure 13.

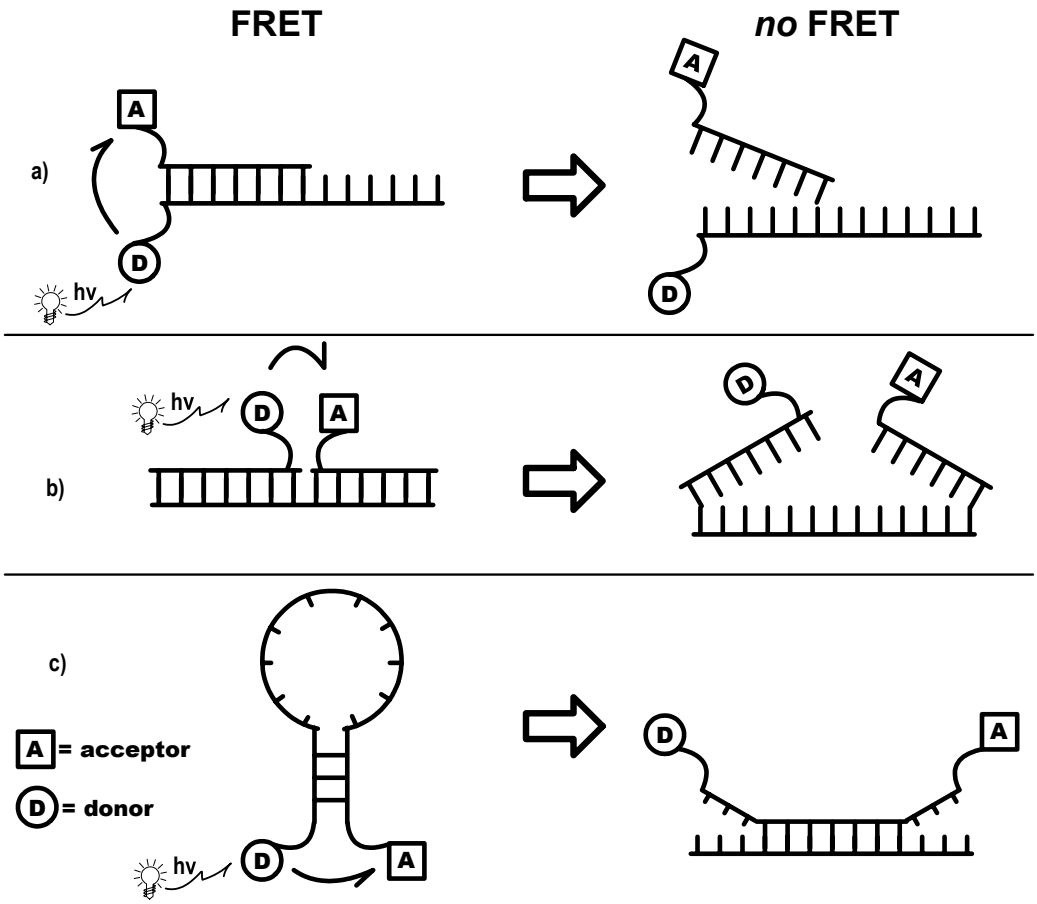


Figure 13: Alternative configurations of FRET donor and acceptor molecules for detecting DNA hybridization. A) The donor and acceptor located on opposite strands is common labeling scheme. B) Attaching the FRET donor and acceptor to two probes that hybridize adjacently is used in such assays as in the lightcycler® assays or in OLA. C) The molecular beacon arrangement involves a double labeled oligo. The ends are complementary forming a stem structure when no target is present. Hybridization of the DNA sequence in the loop structure to a target causes separation of the stem and loss of fluorescence.

*Note: Contrary to common belief, FRET does not involve the emission of light by the donor followed by absorption by the acceptor. That is to say there is no intermediate photon but rather energy transfer occurs via dipole-dipole interaction. For this reason many specialists refer to the process as simply resonance energy transfer or RET.

Chemiluminescence

Chemiluminescence is a special type of *luminescence* in which the emission of light is the result of a chemical or biochemical reaction. Often the process involves the enzymatic conversion of a substrate from one form to another, and the conversion process releases a detectable frequency of light. In the case of Pyrosequencing [143, 144], a cascade of enzymes are used to convert pyrophosphate (a chemical byproduct of the primer-extension reaction) to a reactive ATP which can then participate in the light-producing catalysis of luciferen by luciferase [144, 145].

Fluorescence polarization (FP)

Another fluorescence property that can be exploited for allele detection is how a fluorophore responds to polarized light. If the light from the excitation source is filtered so that the emitted photons travel along the same plane, the fluorescent emission will follow a corresponding angle (polarized emission) depending on a number of factors. Basically, if the dye is immobilized to a long molecule, the deflection angle of incoming and outgoing photons will be more predictable. The extreme opposite condition is when the dye is free in solution and can tumble and rotate easily. In such case, there is much less correlation between incoming and outgoing photon angles (depolarized emission). In genotyping strategies [146], attachment or removal of fluorescently labeled nucleotides can be followed through measuring changing in FP.

Commercially available genotyping systems

Combinations of reaction principle, reaction format, and detection mechanism are many and varied, and a number of full-package genotyping systems are available on the market. Some commercially available genotyping platforms accommodate variations on a theme and thus can be used for alternative strategies of allele-detection. For example, Applied Biosytems (Foster city, CA, www.appliedbiosystems.com) supplies a device called the 7700 sequence detector ® and recently released a high-throughput version called the 7900HT. This was one of the first instruments to combine PCR thermocycling with multi-channel fluorescence detection.

The original SNP genotyping assay supported by this platform is called 5' nuclease or "Taqman" assay [147, 148]. The Taqman concept involves montoring FRET between a donor and acceptor pair that are attached to the same allele-specific probe. Two alternative allele-specific probes (bearing different donor-acceptor pairs) are added in

the PCR reaction mix prior to the start of amplification. Probe length and PCR conditions are adjusted so that only the allele-specific probe with 100% complementarity to the PCR amplified target will be annealed during the extension phase of the PCR. Since the enzyme used for PCR amplification (taq polymerase) also has exonuclease activity, the allele-specific probe that is annealed to the target will be digested releasing the donor and acceptor molecules and thereby destroying FRET. Alleles are detected by monitoring the increase in fluorescence of donor from the digested allele-specific probes.

Alternatively, "molecular beacons" [149] (figure 13c) can be used for allele detection [150-152] using the same device. Hybridization in such case is registered by the annealing of the allele-specific sequence (located in the "loop" structure) to the PCR target which in turn separates the FRET donor and acceptor (located in opposing position in the "stem" structure). The scorpion assays [153] in effect attach a molecular beacon directly to the PCR primer and allele discrimination can either be through allele-specific PCR [154] or allele specific hybridization [155]. The attractive feature of these techniques is that the "closed tube" homogeneous reaction requires no addition or transfer of components for allele-discrimination.

Other platforms offer streamline implementations of specific genotyping strategies. For example, several companies provide SNP genotyping systems centered on *DNA microarrays*. Affymetrix (Santa Clara, CA, www.affymetrix.com) offers the genechip ® technology [18] where high-density sets of up to 10^6 allele-specific probes per cm^2 can be photo-lithographically synthesized [156] on a glass slide. Several multiplex PCRs (with fluorescently labeled primers) serve to amplify and label the polymorphic loci, and once combined and concentrated, the complex pool is hybridized to the genechip. Hybridization conditions are adjusted so as to optimize allele-specific hybridization to the immobilized probes, and alleles are inferred by fluorescent imaging. Nanogen (San Diego, CA, www.nanogen.com) employs electrical fields to direct DNA molecules to specific locations on a chip [157] and allele distinction is carried out through gradually reversing the polarity of the electrical field [158] and examining the melting patterns of the probe/target duplexes at the microelectrodes. Alternatively, as in array primer extension or APEX (Asper Biotech Ltd., Tartu, Estonia, www.asperbio.com), PCR amplified targets are hybridized to pre-arrayed SNP specific probes and alleles are called by determining which type of nucleotide was incorporated during a single-base extension (SBE) reaction [159]. A fluorescence-based oligo ligation assay (OLA) has also been demonstrated on microarrays [160].

Instead of planar arrangements, two companies have pioneered the creation of coded microsphere or "bead" arrays. The approach taken by Luminex (Austin, TX, www.luminexcorp.com) is to capture allele-specific products from OLA [161] or SBE [162,163] reactions to the beads, and decipher both bead and genotype identity through a flow cytometer. Illumina (La Jolla, CA, www.illumina.com) also uses similar bead chemistries but fluorescence detection takes place by trapping the beads onto a fiber-optic bundle array [164].

For Mass-spec genotyping, Sequenom (San Diego, CA, www.sequenom.com) has coupled solid-phase purification into their MassArray® and more recently the MassEXTEND® systems. In the PROBE ™ assay, SBE products are purified through captured and cleaning on magnetic particles. In the PinPoint ™ (Applied Biosystems) assay, the purification takes place on reverse-phase material tips called ZipTips (Millipore, Bedford, MA, www.millipore.com). Another single-tube purification and preparation system for SBE reactions is the GOOD Assay [136,165]. The protocol serves to both remove reagents that can interfere with interpretation of MS results and reduce the size of the SBE allele-specific products to optimal lengths for MS analysis. Finally, Qiagen (Bothell, WA, www.qiagen.com) offers a MassCode ™ system for SBE genotyping by MALDI-TOF MS. The system enrolls MassTags or special labels that, when attached, can identify allele-specific PCR products by their distinct mass signal in MS analysis.

The Invader® Assay (Third Wave Technologies, Maddison, WI, www.twt.com) employs a special digestive enzyme called a FLAP endonuclease (FEN) [166] for the allele-discrimination reaction. The enzyme recognizes and selectively cleaves a structure formed by the co-hybridization of two overlapping oligonucleotide probes. If the invading allele-specific probe matches the target at the junction, the enzyme cleaves off the overhanging portion of the downstream probe. For signal amplification, the cleaved portion of downstream probe can be used as the invading probe for a secondary generic Invader reaction [167]. Allele-specific products have been measured using MS [168], FRET [167] or FP [122, 169]. The yet-to-be-fully-realized potential with this technique was that it can be performed on genomic DNA (i.e., does not require PCR) and it is a closed tube reaction.

Another genotyping platform based on SBE is the SNPit or SNPstream™ systems provided by Orchid Biosciences (Princeton, NJ, www.orchid.com). Allele-specific products are captured onto streptavidin-coated microtiter plates or, as in the recently advertised SNPit tag array ™, to an addressable microarray and the labels are

detected through an elisa-like colorimetric reaction [170]. Also, the SNaPshot ™ system (Applied Biosystems) provides a SBE genotyping strategy that can be run on either slabgel or capillary sequencing machines [159]. Separation of alleles in this strategy can be enhanced by the addition of 5' tails of varying length to the SBE primers [171].

Additional gel-based genotyping platforms include the multiplex allele-specific diagnostic assay (MASDA) [172] based on allele-specific hybridization resulting in sequence specific band patterns upon electrophoretic separation. Another strategy called microplate array diagonal gel electrophoresis (MADGE) [173] accomplishes genotyping through fluorescent detection of allele-specific PCR reactions or digests. The oligo ligation assay (OLA) [174] has been adopted for SNP genotyping on gels [175]. In practice, OLA has proven compatible with a wide-variety of reaction formats and detection mechanisms. For instance, OLA combined with colorimetric detection has been also been performed on microtiter plates [176], while fluorescent labeled OLA has been detected using DNA sequencing equipment [177]. An emerging genotyping company (ParAllele biosciences,San Francisco, CA, www.p-gene.com) has recently started service genotyping based on OLA and padlock probes [192, 178, 179] and rolling circle amplification (RCA) [180].

The descriptions given here are admitantly very brief so interested readers are advised to explore the primary publications for deeper and a more comprehensive description of each of these innovative genotyping technologies and platforms.

ETHICS AND GENETIC RESEARCH

"We are not pushing back the hands of time, but rather pushing them forward".

WMH

The scientific accomplishments of genetic research during the 20 years have thundered into mainstream society. Accomplishments are so quick and breakthroughs so frequent and unpredictable that the ethical repercussions often follow as an echo or aftershock. As the dust settles, many fascinating and challenging issues are exposed to consideration.

For example, there is no question that medical progress has extended the average human lifespan, but what do these extra years bring? I would argue that at present we are not making the average person more youthful, but rather making old people older. Amidst the medical miracles that stave off infection and medicines that compensate for physiological deficiencies, it is important to consider the research we can do to allow people to appreciate the added "golden years". Along this vein, research into late-onset disorders such as Alzheimer's disease is hoped to improve the *quality* of life to match the extended *quantity* of life afforded by modern medicine.

Similarly, the paradigm of scientific reductionism drives us to collect and organize the sequence of the human genome with careful attention to detail. The expectation is that clarifying our "genetic-identity" will lead to insight into ourselves as a whole. The key to understanding the similarities and differences that we have as a race is assumed to exist in the subtle variation in our genetic code. We are at a vulnerable stage, however, since the advanced technology to collect the information is much more developed than our scientific ability to interpret it.

Participation in modern genetic research thus allows the opportunity and responsibility to consider what will happen to the information we gather from the science we are performing now. We influence the direction science will take tomorrow by choosing what to explore today. Some people attempt to shield academic science as being immune to ethical consideration because, after all, the controversy lies in the application rather than the pure pursuit of knowledge. I tend to agree with another modern day scientist and philosopher who wrote "*Technology must not outweigh our humanity*" – Albert Einstein. It is better to educate when possible, warn when necessary, and in that way make it so we can all sleep better at night.

PRESENT INVESTIGATIONS

Aims

Papers I & II
To test a number of candidate SNPs for association in sporadic early onset Alzheimer's disease.

Papers III, IV & V
To introduce and advance the technique Dynamic Allele Specific Hybridization, a method for genotyping single nucleotide polymorphisms.

PAPERS I & II – ASSOCIATION STUDIES IN ALZHEIMER'S DISEASE

Conceptual overview

Publications I and II both investigate a series of polymorphisms for association in Alzheimer's disease (AD). Paper I [181] presents analysis of polymorphisms in the APOE, VLDL-R, PSEN1, and PSEN2 genes and several statistically significant associations were detected. Paper II presents a through follow-up study one of these signals (in PSEN2) using an expanded set SNP markers positioned throughout the gene.

Experimental summary

The DNA samples used in both papers I and II consisted of patients and controls from the Lothian region of Scotland and were received through collaboration with two different groups [182,183]. The genetic variants studied in paper I included previously reported AD associated polymorphisms in the APOE [184], PSEN1 [185] and the VLDL-R [186] genes. The PSEN2 polymorphisms studied in both papers I and II were identified through SNP discovery efforts by our group [187].

A wide variety of traditional genotyping techniques were used to score the polymorphisms in paper I. Alleles for both the APOE and PSEN1 SNPs were determined using PCR/restriction digestion followed by electrophoretic separation on an agarose gel. The repeat lengths from VLDL-R microsatellite polymorphism were determined using standard microsatellite genotyping procedures. The PSEN2 polymorphism was analyzed via the "dotblot" technique [188]. For paper II, most of the genotyping data was generated using dynamic allele-specific hybridization (DASH) which is fully described in the overview of papers III, IV, and V.

Major results

As presented in paper I, individuals with one or more APOE-ε4 alleles were much more likely to be in our collection of cases than in controls. This was expected, as APOE-ε4 / AD appears to be one of the rare associations that is possible to replicate in nearly all populations around the world [189]. We did not, however, find association with the other previously published candidate alleles, namely the reported 5-unit repeat of the VLDL-R microsatellite or with homozygosity of the "1" allele for PSEN1. Instead, *absence* of the 8-unit repeat of the VLDL-R as well as homozygosity of the "2" allele

of PSEN1, and homozygosity of the "C" allele of PSEN2 were all statistically significant observations in our study.

Since the number of samples included in the initial study was low, and the fact that we had started to develop a fairly dense map of SNPs in the PSEN2 gene, we decided to extend the PSEN2 association analysis in a new study including a second set of Scottish EOAD samples. The results as presented in paper II are of 8 SNPs located both upstream and downstream of the SNP locus that gave the original association in Paper I. In the new study, no association was detected for any of the loci using a standard χ^2 test. When we stratified the material based on the presence of one or more APOE-ε4 alleles, three polymorphisms showed an association where the p-value reached ~0.05. Alleles of these three SNPs were found to fall into just 2 haplotypes, and one SNP locus was chosen to function as an htSNP [70] to test in the second independent set of AD cases and controls. Chi-square analysis of this genotyping data did not replicate the positive association found in the first set AD cases and controls. Although no association study can completely rule out the involvement of gene in a disease, the ultimate result of papers I & II is that no supportive evidence was generated asserting the involvement of the PSEN2 gene with sporadic early-onset Alzheimer's disease.

PAPERS III, IV, V – DYNAMIC ALLELE-SPECIFIC HYBRIDIZATION

Introduction to the method

Dynamic allele-specific hybridization (DASH) was created as a method to genotype SNPs as well as short indel polymorphisms. The three basic concepts on which the technology is built are:

- **Reaction Principle** Hybridization
- **Reaction Format** Solid-phase
- **Detection mechanism** Fluorescent detection

Since the original publication [190], a number of refinements have been made in regard to **reaction format** and **detection mechanism**, however the core assay procedure remains as follows:

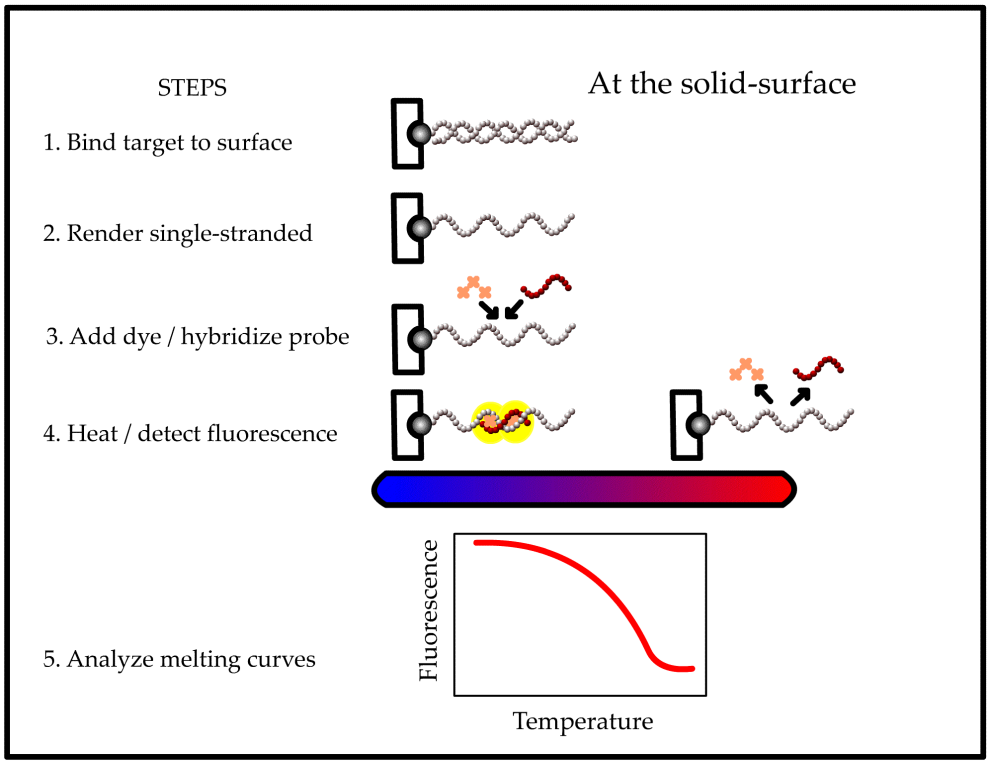


Figure 14: Core reaction steps in the DASH assay procedure. 1) After PCR amplification of the target SNP, the biotinylated PCR products are immobilized to a streptavidin-coated surface. 2) Rinsing the samples with an alkali solution serves to both denature the PCR product and remove non-biotinylated strand along with other PCR byproducts. 3) A probe, specific for one of the SNP alleles, is hybridized along with SYBR Green I dye. 4) The sample is heated while simultaneously monitoring fluorescence. 5) The collected fluorescence vs. temperature data is plotted, and genotypes are scored through comparative analysis of the generated melting curves.

The initial format involved capture of the PCR products to streptavidin-coated microtiter plate wells. Using this solid-phase format was very convenient as all preparation steps thereafter involved either simple addition or removal of solutions to or from the wells. Even detection, as first performed using the ABI 7700 sequence detector, took place using the same microtiter plate. The original format thus allowed for the genotyping of 96 separate DNAs in a single run.

The dye used for fluorescence detection, as described in paper III, was Sybr Green I. The characteristic that made this dye suitable for detecting the melting of DNA duplexes was 10,000 fold increase in fluorescence generated by the dye binding to double-stranded DNA. Thus, when the probe is annealed to the immobilized target, the dye fluoresces brightly. When the samples are heated up and the *melting temperature (Tm)* is reached, the dye will be released and stop fluorescing. The large drop in fluorescence is thus an indication of the sample reaching the Tm for the probe/target duplex. The Tm is then used to classify which SNP alleles are present in the sample. The easiest way to visualize the drop in fluorescence is to plot the fluorescence vs. temperature values collected during the heating step. Further clarity of the melting event is possible through plotting the negative derivative of the fluorescence data as the Tm is then indicated by the presence of a peak (figure 15).

Melting Temperature (Tm):

Under defined assay conditions, the Tm is the temperature at which the DNA strands of a DNA duplex will separate.

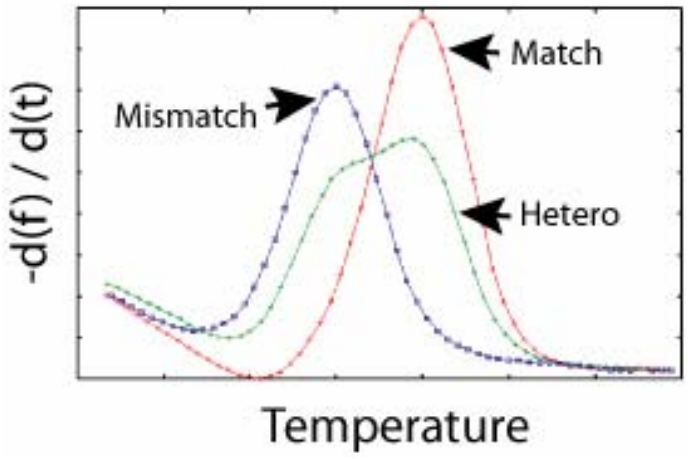


Figure 15: Results from a DASH assay for three samples. Each sample has a different genotype representing the three possible genotypes from an SNP position. A "match" peak represents a target that is complementary to the probe, a "mismatch" peak is non-complementary to the probe, and a "heterozygous" sample is indicated by the presence of both types of peaks.

In paper III, DASH is demonstrated to function on all potential pairings and mispairings of bases that can arise from hybridization of the allele-specific probe to the complement or non-complement target. In addition, the DASH strategy that would later be commercialized by Thermo-Hybaid (www.thermohybaid.com) is outlined in detail. The device, now termed the melting curve analysis (MCA) device, is currently available from Thermo-Hybaid along with all the necessary accessories for performing the original version of DASH.

iFRET – Improving fluorescence detection

Fluorescent detection using only Sybr Green I has a number of advantages. The dye is very inexpensive and is also easy to detect due to the large amount of fluorescence it produces. The main drawback of dye has to do with background fluorescence. Since the Sybr Green will fluoresce when bound to any double-strand segment of DNA, secondary structures, such as those caused by the immobilized target folding back on itself, can interfere with detection of DASH melting curves [130].

An alternative strategy for fluorescent detection of DNA hybridization is through fluorescent resonant energy transfer (FRET) which is fully described earlier in this thesis. The advantage of this detection system is that FRET occurs only when the donor and acceptor molecules are kept close together. When the dyes are separated, energy transfer is no longer possible. An increase in fluorescence from the donor and a decrease fluorescence of the acceptor is thus indicative of DNA duplex denaturation. Alternative spatial relationships of FRET donor and acceptor molecules are depicted in figure 13 of this thesis. The drawback with FRET, however, is the high cost of fluorescently labeling DNA and the relatively low yield of fluorescence.

Paper IV introduces a new alternative for fluorescent detection of hybridization that maintains the high fluorescence as observed with Sybr Green detection, but combines the specificity for particular DNA duplexes that is normally exclusively reserved for FRET detection. The innovation behind *induced fluorescent resonance energy transfer* or *iFRET*, is the use of Sybr Green I as a donor molecule for a FRET reaction between this intercalating dye and an acceptor attached to the allele-specific probe (figure 16). Since the fluorescence from the acceptor molecule is maximized far down in the visible spectrum, fluorescent emission from the acceptor can easily be discerned from fluorescence given off by Sybr Green alone. Thus it is possible to effectively filter out background fluorescence caused by the non-specific nature of Sybr Green I.

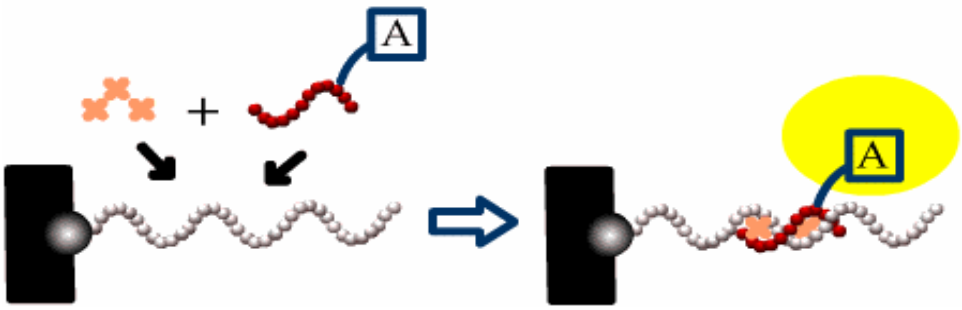


Figure 16: The reaction principle behind induced fluorescent resonance energy transfer (iFRET). An allele-specific probe labeled with a FRET acceptor molecule is hybridized in the presence of Sybr Green I dye.

The efficiency of energy transfer between Sybr Green I and the common FRET acceptor called 6-Rhodamine or ROX, is quite striking. As depicted in figure 17, the ROX acceptor is able to harvest the energy from Sybr Green I with high effectiveness. In addition, the emission spectrum for ROX is much narrower than that of Sybr Green I which in turn leads to a more "condensed" fluorescent signal. This can explain the relatively stronger fluorescence signal observed in paper IV when comparing Sybr Green I to iFRET emissions.

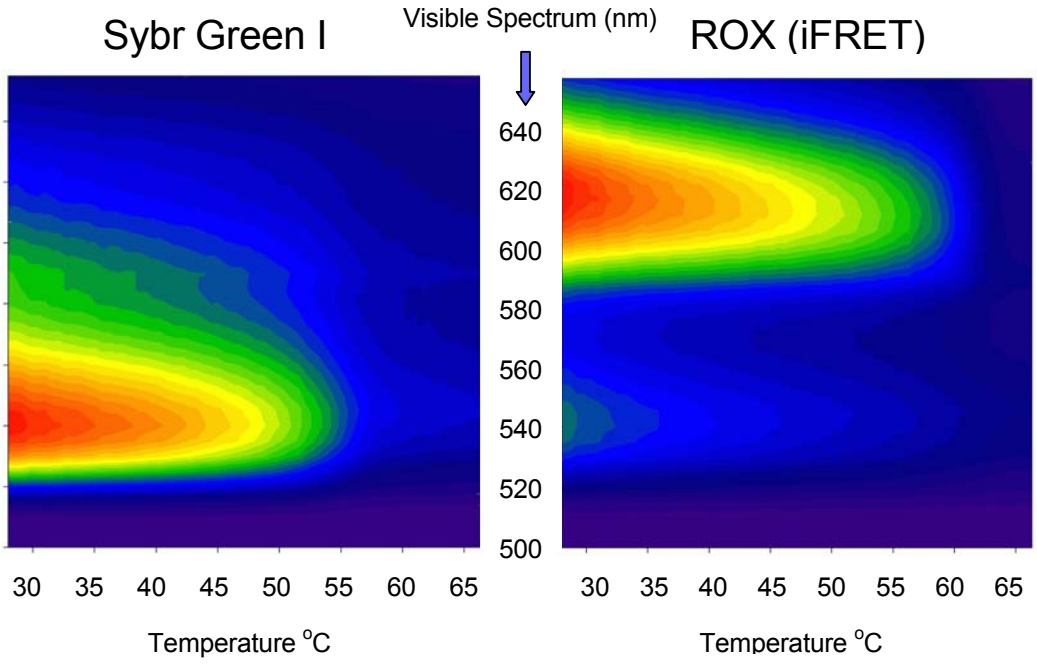


Figure 17: DASH runs comparing Sybr Green I and iFRET detection. In contrast to normal DASH, the depiction of the fluorescence from the entire visible spectrum is plotted throughout the entire run. Fluorescence intensity is indicated by the color range, with red being maximum (100%) and blue being minimum (0% or background).

DASH-2: Macro-array format

In response to pressure from the scientific community to increase genotyping throughput and decrease costs, paper V presents a new version of DASH. A major step forward was the conversion of the DASH assay to an array format. To make the system back compatible with DASH-1 on microtiter plates, the immobilization chemistry used for the new array format was still biotin-streptavidin interaction. Two alternative approaches for creating the arrays are compatible with the new DASH-2 system. The most straight-forward solution and least hardware-dependent is through centrifugation [191]. In such case, the streptavidin-coated membrane is clamped to microtiter plate used to generate the PCR, and PCR products are transferred to the membrane by simple centrifugation. The spin array / DASH-2 combination has been tested and confirmed when transferring PCR products from 96, 384, and even 1536 well microtiter plates. Alternatively, robotic spotting has also been used to create arrays. The advantage with robotics is the possibility to create arrays with higher density of features.

Flexibility and multiplex options

The implementation of arrays, together with iFRET detection, offers a number of opportunities for the user to customize the DASH-2 system to their specific needs. The scale of the experiment can be adjusted by spin-transferring the PCR products from microtiter plates containing a higher or lower number of wells. Robotic spotting can be used in case of ultra-high throughput requirements. Another way to increase the number of genotypes is through multiplexing. As demonstrated in paper V, there are several possible multiplexing options. For example, array features can be made to contain PCR products from more than one SNP locus either through pooling or multiplex PCR. Genotypes for each SNP in the mixed feature can be determined using spectral multiplexing (involving iFRET and different acceptors for each SNP locus) or serially (involving reprocessing of the membrane). If few samples but many different SNPs are to be studied, an array can be created that has features from numerous SNPs. Subsequently a "probe-cocktail" or mixture of the probes for all the SNPs to be interrogated can be hybridized to the membrane at one time. Up to 250 probes were combined in this matter and still resulted in readable genotypes for each sample.

Prototyping the DASH-2 system

One of the more challenging aspects of the current research has been the actual construction of the DASH-2 device itself. No commercial products were available for heating the microtiter-plate size arrays and detecting fluorescence, so we set out to assemble and integrate all the necessary components from scratch. An overview of the current system is depicted in figure 18.

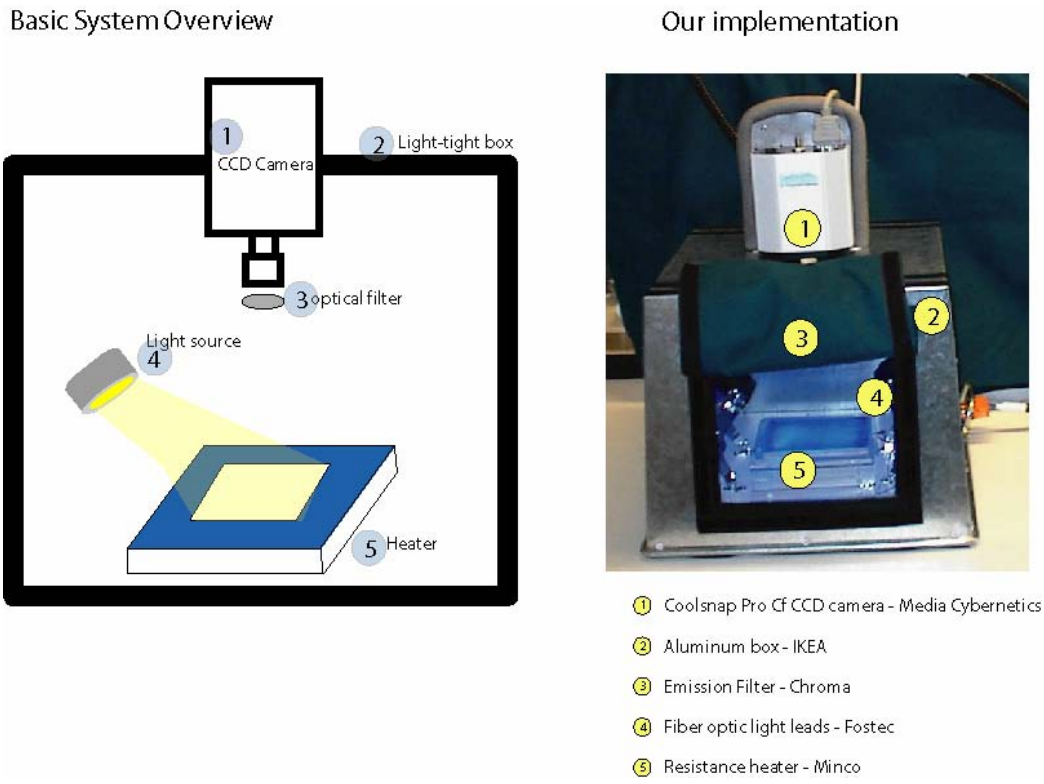


Figure 18: The essential components of the DASH-2 detection system depicted both in cartoon and realized form.

Creation of a fully functional system involved the construction of a variety of components from their basic parts. For example, the heating system (required to raise the temperature of the array at a constant rate) was created using a thin resistance heater commonly used in satellites to keep panels warm. Integration of the heat controller with temperature sensors involved additional components depicted in figure 19. Additional engineering solutions for

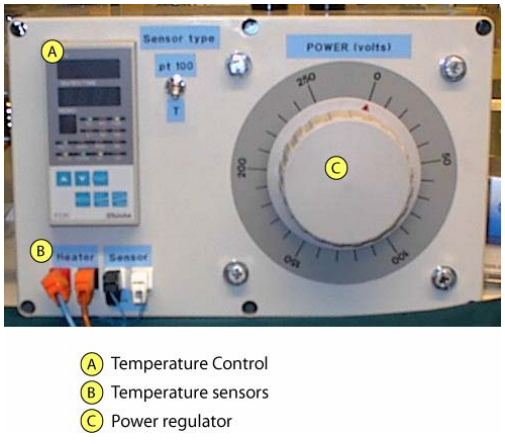


Figure 19: Additional equipment required to control temperature during the DASH-2 procedure.

array manufacture and fluidic processing were also incorporated into the final functional system.

Conclusion

The main objective of paper V was to demonstrate features of the DASH-2 system that make it an attractive alternative for SNP genotyping. The improvements to the original method have been to increase flexibility and throughput without compromising quality or incurring extra costs. The alternative approaches for array creation coupled with spectral, serial, and/or inter-feature multiplexing offer the opportunity for performing SNP genotyping at several different scales depending on the design of the study.

ACKNOWLEDGEMENTS

A number of people have been, and continue to be, instrumental to the fulfillment graduate education. My most sincere gratitude goes to:

Dr. Anthony Brookes: A true believer in me and the formula "1x Inspiration + 50x Perspiration = Dissertation". You are a dedicated driving force in science. I wish you success and happiness in all your future endeavors.
Dr. Hans Wigsell: for bits of wisdom regarding our first company, and reminding me that the key to success is to "keep it fun".
Dr. Claes Wahlestedt: for bringing our group to the Center for Genome Research, and providing the invisible hand of administrative support when it was needed.
The **three Ulf's** (**Dr. Ulf Gyllensten**, **Dr. Ulf Landegren**, and **Dr. Ulf Petterson**): each of which a unique cornerstone of the Uppsala University department formerly known as MedGen. Thanks to each of you for taking time to assist me on the occasions when it was necessary.
Dr. Joe Terwillger: For constantly reminding me of the other side of the coin, and for joining the fishing trip in which the guide couldn't even find the water.
Dr. Carl Johan Sundberg: For interesting discussions about the public dissemination of science, and the eagerness to start entrepreneurial activities at KI. Also, for help with curriculum details.
Dr. Mikael Holst: for your compassion and support with many the administration details.

All members, past and present, of the Brookes group:
Dr. Magnus Jobs: an artist in the genetic arts, and my scientific brother-in-arms. The opposing bookend that held the DASH development projects together. Of all the things I've gained from these doctoral years, your friendship is my most valued. **Dr. Lars Feuk:** I thank you for deep scientific discussions, the shallow ones too, for football and horses, and for all the grape advice. At times you have been a keen adversary, other times a role model, but always a trusted companion. **Dr. Testfight Emahazion**: for teaching me that this education is a finite task and the advice to invest in FramFab. **Dr. Jonathan Prince**: for your friendship and the inspiration to publish. **Sarah Sawyer:** for helping maintain perspective, and understanding my roots. I know there is the right "fleece-factor" somewhere for you. **David PANDA FISH Fredman**: for computer expertise and adding color to the group. **Dr. Torsten Mayr & Melanie**: for adding the chemical dimension to the project. **Linda Strömqvist** and **Alex Kovacs**: for their tireless efforts in the DASH lab. Where would we be without you? **Daniel Pedersson** and **Johan Klingberg:** for their keen digital support and trips to Volcano. Again, where would we be without you? **Gillian Munns and Daniel Rios:** for taking time for conversations. **Marianne Siegfried**: for the knupfli dinner, and efforts to make the group better. **Dr. Harvest Gu:** for great talks in Iceland and enthusiasm for quality research. **Dr. Salim Mottagui-Tabar**: for friendship and a tour of the internet.
and the rest of the staff and technical crew **Suzzane Potter, Astrid Lindstedt, Camilla Lagerberg, Sandra Tammer, Nina Junttila, Hagit Katzov, Vivian Daza-Saucedo, Linda Berglind, Per Ewing, Satu Valtonen**: for putting up with all the beta-testing, odd experiments, and thousands of DASH runs.

The CGB family at Karolinska Institute:
Especially **Dr. Björn Andersson** and **Dr. Martti Tammi** for their friendship, advice, and eagerness to join in spirited conversations. **Kim Andersson**: for great parties.
The MedGen folks:
Dr. Patrik Magnusson: for -25°C skiing, night skating, and for dragging my new laptop over the Atlantic. Who would have imagined that Minna and Liv would arrive just hours and meters apart! **Dr. Tomas Bergström & Dr. Lucia Cavalier** for their unwavering support, friendship through my early years at MedGen. **Hasse Engkvist**: for sailing and skating and always having the answers to even the strangest question I could come up with. **Per-Ivan Wyoni**: for fishing, skating, good talks, and the technical computer help. **Dr. Anna Beskow**: for frank and open kindness. **Dr. Agnetta Josefsson**: for practical help with the taqman and friendship. **Dr. Marta Alarcón & Dr. Anna-Karin Lindqvist**: for teaching me the ropes early on. **Dr. Veronica Magnusson, Cecilia Johansson, Martin Moberg, & Paula Jalonene:** for coffee room discussions. **Dr. Bo Johanneson**: for being my study partner. Thanks for the tips on thesis too! **Dr. Mats Nilsson:** for the sharp, innovative enthusiasm for science. **Dr. Eva Lindholm:** for keeping me in the loop during the mitochondrial year. **Dr. Elena Jazin:** for taking me under her wing in the very beginning. **Inger Eriksson and Inger Jonasson**: for keeping the lab running. The backbone of medical genetics, **Jeanette Backman, Elisabeth Sandberg, and Annette Uselius:** for the administrative magic that made the seemingly impossible possible. **Stig Söderberg**: for access to all the tools, and a guiding hand. **Eddie Lundh**: for the practical help.
Special thanks to **Gregor Blomquist (MTC) & Åke Samuelsson (CLC Products AB)**: For access and education to the practical components that made many of the prototyping efforts possible. **Molecular Probes** and **Claes Ohlson**:for selling the coolest stuff, and **Thermo-Hybaid**: for taking a chance and commercializing the first DASH device. Special thanks to crew at Ashford.

Fellow scientific colleagues from afar: **Dr. Jessie Theuns, Dr. Jurgen Del-Favero, Tanja Näslund, Dr. George Mellick, Dr. Dave Austin, Ryan & Åsa Preston, Cynthia Shuman, Carl Larcina, Dr. Ipek & Dr. Nadir Çiray, Dr. Rudiger Kolm & Dr. Irene Brodowsky, and Jonathan Severn:** for keen perspectives on science and friendship.

Bo Kumlin: for the tips in business and plastic padding, and **Birgit Kumlin:** for the numerous meals along the way. To my Swedish family, **Elsa-Britt**, **Mia**, and **Ninni** for the many midsummer memories and always keeping a door open when I needed shelter.

My mother and father for their encouragement and support through these long years in Sweden. To Brothers **Jim** and **Sean** for volleyball, friendship, and keeping the family contact alive across the Atlantic. Sister **René** for the genuine concern, and boosts of moral in my times of need. Sister **Pam** for being the example of success.

And to my own family, **Åsa Kumlin**, my rock-steady support through these roller coaster graduate study years. Thank you for your relentless patience, ceaseless pep-talks, and your unwavering love. And to little **Liv** who decided to appear on the disputation guest list all on her own.

REFERENCES

1. Ebersberger, I., Metzler, D., Schwarz, C. & Paabo, S. **Genomewide comparison of DNA sequences between humans and chimpanzees**. *Am J Hum Genet* **70**, 1490-7. (2002).
2. Cairns, J. **Mutation selection and the natural history of cancer**. *Nature* **255**, 197-200. (1975).
3. Hsieh, P. **Molecular mechanisms of DNA mismatch repair**. *Mutat Res* **486**, 71-87. (2001).
4. Nachman, M. W. & Crowell, S. L. **Estimate of the mutation rate per nucleotide in humans**. *Genetics* **156**, 297-304. (2000).
5. Batzer, M. A. & Deininger, P. L. **Alu repeats and human genomic diversity**. *Nat Rev Genet* **3**, 370-9. (2002).
6. Shi, Y. R. et al. **Genetic analysis of chromosome 22q11.2 markers in congenital heart disease**. *J Clin Lab Anal* **17**, 28-35 (2003).
7. Weber, J. L. & Broman, K. W. **Genotyping for human whole-genome scans: past, present, and future**. *Adv Genet* **42**, 77-96 (2001).
8. Goldstein, D. B. & Schlotterer, C. (Oxford Univ. Press, Oxford, 1999).
9. Kogelnik, A. M., Lott, M. T., Brown, M. D., Navathe, S. B. & Wallace, D. C. **MITOMAP: a human mitochondrial genome database--1998 update**. *Nucleic Acids Res* **26**, 112-5. (1998).
10. Kimura, M. **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences**. *J Mol Evol* **16**, 111-20. (1980).
11. Holliday, R. & Grigg, G. W. **DNA methylation and mutation**. *Mutat Res* **285**, 61-7. (1993).
12. Brookes, A. J. **The essence of SNPs**. *Gene* **234**, 177-86. (1999).
13. Sachidanandam, R. et al. **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms**. *Nature* **409**, 928-33. (2001).
14. Collins, F. S. et al. **New goals for the U.S. Human Genome Project: 1998-2003**. *Science* **282**, 682-9. (1998).
15. Masood, E. **As consortium plans free SNP map of human genome**. *Nature* **398**, 545-6. (1999).
16. Kruglyak, L. & Nickerson, D. A. **Variation is the spice of life**. *Nat Genet* **27**, 234-6. (2001).
17. Cargill, M. et al. **Characterization of single-nucleotide polymorphisms in coding regions of human genes**. *Nat Genet* **22**, 231-8. (1999).
18. Wang, D. G. et al. **Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome**. *Science* **280**, 1077-82. (1998).
19. Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L. & Kwok, P. Y. **Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms**. *Genome Res* **8**, 748-54. (1998).
20. Nickerson, D. A. et al. **DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene**. *Nat Genet* **19**, 233-40. (1998).
21. Halushka, M. K. et al. **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis**. *Nat Genet* **22**, 239-47. (1999).
22. Altshuler, D. et al. **An SNP map of the human genome generated by reduced representation shotgun sequencing**. *Nature* **407**, 513-6. (2000).

23. Lercher, M. J. & Hurst, L. D. **Human SNP variability and mutation rate are higher in regions of high recombination**. *Trends Genet* **18**, 337-40. (2002).
24. Li, W. H. & Sadler, L. A. **Low nucleotide diversity in man**. *Genetics* **129**, 513-23. (1991).
25. Guillaudeux, T., Janer, M., Wong, G. K., Spies, T. & Geraghty, D. E. **The complete genomic sequence of 424,015 bp at the centromeric end of the HLA class I region: gene content and polymorphism**. *Proc Natl Acad Sci U S A* **95**, 9494-9. (1998).
26. Blasco, M. A., Gasser, S. M. & Lingner, J. **Telomeres and telomerase**. *Genes Dev* **13**, 2353-9. (1999).
27. Henikoff, S., Ahmad, K. & Malik, H. S. **The centromere paradox: stable inheritance with rapidly evolving DNA**. *Science* **293**, 1098-102. (2001).
28. Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. **Introducing RefSeq and LocusLink: curated human genome resources at the NCBI**. *Trends Genet* **16**, 44-7. (2000).
29. Small, K. M., Seman, C. A., Castator, A., Brown, K. M. & Liggett, S. B. **False positive non-synonymous polymorphisms of G-protein coupled receptor genes**. *FEBS Lett* **516**, 253-6. (2002).
30. Gabriel, S. B. et al. **The structure of haplotype blocks in the human genome**. *Science* **296**, 2225-9. (2002).
31. Goddard, K. A., Hopkins, P. J., Hall, J. M. & Witte, J. S. **Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations**. *Am J Hum Genet* **66**, 216-34. (2000).
32. Barbujani, G., Magagni, A., Minch, E. & Cavalli-Sforza, L. L. **An apportionment of human DNA diversity**. *Proc Natl Acad Sci U S A* **94**, 4516-9. (1997).
33. Yu, A. et al. **Comparison of human genetic and sequence-based physical maps**. *Nature* **409**, 951-3. (2001).
34. Dunham, I. et al. **The DNA sequence of human chromosome 22**. *Nature* **402**, 489-95. (1999).
35. Hattori, M. et al. **The DNA sequence of human chromosome 21**. *Nature* **405**, 311-9. (2000).
36. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. **Comprehensive human genetic maps: individual and sex-specific variation in recombination**. *Am J Hum Genet* **63**, 861-9. (1998).
37. Weiss, K. M. & Clark, A. G. **Linkage disequilibrium and the mapping of complex human traits**. *Trends Genet* **18**, 19-24. (2002).
38. Laan, M. & Paabo, S. **Demographic history and linkage disequilibrium in human populations**. *Nat Genet* **17**, 435-8. (1997).
39. Slatkin, M. **Disequilibrium mapping of a quantitative-trait locus in an expanding population**. *Am J Hum Genet* **64**, 1764-72. (1999).
40. Risch, N. **Linkage strategies for genetically complex traits. II. The power of affected relative pairs**. *Am J Hum Genet* **46**, 229-41. (1990).
41. MacGregor, A. J., Snieder, H., Schork, N. J. & Spector, T. D. **Twins. Novel uses to study complex traits and genetic diseases**. *Trends Genet* **16**, 131-4. (2000).
42. Lander, E. S. & Schork, N. J. **Genetic dissection of complex traits**. *Science* **265**, 2037-48. (1994).
43. Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. **A comprehensive review of genetic association studies**. *Genet Med* **4**, 45-61. (2002).
44. Gatz, M. et al. **Heritability for Alzheimer's disease: the study of dementia in Swedish twins**. *J Gerontol A Biol Sci Med Sci* **52**, M117-25. (1997).

45. Terwilliger, J. D., Haghighi, F., Hiekkalinna, T. S. & Goring, H. H. **A bias-ed assessment of the use of SNPs in human complex traits**. *Curr Opin Genet Dev* **12**, 726-34. (2002).
46. Schulze, T. G. & McMahon, F. J. **Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines**. *Am J Med Genet* **114**, 1-11. (2002).
47. Saunders, A. M. et al. **Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease**. *Neurology* **43**, 1467-72. (1993).
48. Hugot, J. P. et al. **Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease**. *Nature* **411**, 599-603. (2001).
49. Vignal, A., Milan, D., SanCristobal, M. & Eggen, A. **A review on SNP and other types of molecular markers and their use in animal genetics**. *Genet Sel Evol* **34**, 275-305. (2002).
50. Li, W. H., Gojobori, T. & Nei, M. **Pseudogenes as a paradigm of neutral evolution**. *Nature* **292**, 237-9. (1981).
51. Martinez-Arias, R. et al. **Sequence variability of a human pseudogene**. *Genome Res* **11**, 1071-85. (2001).
52. Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. **Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations**. *Proc Natl Acad Sci U S A* **95**, 10774-8. (1998).
53. Risch, N. J. **Searching for genetic determinants in the new millennium**. *Nature* **405**, 847-56. (2000).
54. Lander, E. S. **The new genomics: global views of biology**. *Science* **274**, 536-9. (1996).
55. Chakravarti, A. **Population genetics--making sense out of sequence**. *Nat Genet* **21**, 56-60. (1999).
56. Pritchard, J. K. **Are rare variants responsible for susceptibility to complex diseases?** *Am J Hum Genet* **69**, 124-37. (2001).
57. Slager, S. L., Huang, J. & Vieland, V. J. **Effect of allelic heterogeneity on the power of the transmission disequilibrium test**. *Genet Epidemiol* **18**, 143-56. (2000).
58. Reich, D. E. & Lander, E. S. **On the allelic spectrum of human disease**. *Trends Genet* **17**, 502-10. (2001).
59. Pritchard, J. K. & Cox, N. J. **The allelic architecture of human disease genes: common disease-common variant...or not?** *Hum Mol Genet* **11**, 2417-23. (2002).
60. Risch, N. & Merikangas, K. **The future of genetic studies of complex human diseases**. *Science* **273**, 1516-7. (1996).
61. Kruglyak, L. **What is significant in whole-genome linkage disequilibrium studies?** *Am J Hum Genet* **61**, 810-2. (1997).
62. Collins, A., Lonjou, C. & Morton, N. E. **Genetic epidemiology of single-nucleotide polymorphisms**. *Proc Natl Acad Sci U S A* **96**, 15173-7. (1999).
63. Stephens, J. C. et al. **Haplotype variation and linkage disequilibrium in 313 human genes**. *Science* **293**, 489-93. (2001).
64. Patil, N. et al. **Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21**. *Science* **294**, 1719-23. (2001).
65. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. **High-resolution haplotype structure in the human genome**. *Nat Genet* **29**, 229-32. (2001).

66. Judson, R., Salisbury, B., Schneider, J., Windemuth, A. & Stephens, J. C. **How many SNPs does a genome-wide haplotype map require?** *Pharmacogenomics* **3**, 379-91. (2002).
67. Kruglyak, L. **Prospects for whole-genome linkage disequilibrium mapping of common disease genes**. *Nat Genet* **22**, 139-44. (1999).
68. Reich, D. E. et al. **Linkage disequilibrium in the human genome**. *Nature* **411**, 199-204. (2001).
69. Dawson, E. et al. **A first-generation linkage disequilibrium map of human chromosome 22**. *Nature* **418**, 544-8. (2002).
70. Johnson, G. C. et al. **Haplotype tagging for the identification of common disease genes**. *Nat Genet* **29**, 233-7. (2001).
71. Devlin, B. & Risch, N. **A comparison of linkage disequilibrium measures for fine-scale mapping**. *Genomics* **29**, 311-22. (1995).
72. Pritchard, J. K. & Przeworski, M. **Linkage disequilibrium in humans: models and data**. *Am J Hum Genet* **69**, 1-14. (2001).
73. Olivier, M. et al. **Complex high-resolution linkage disequilibrium and haplotype patterns of single-nucleotide polymorphisms in 2.5 Mb of sequence on human chromosome 21**. *Genomics* **78**, 64-72. (2001).
74. Tishkoff, S. A. et al. **A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations**. *Am J Hum Genet* **62**, 1389-402. (1998).
75. Kidd, K. K. et al. **A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus**. *Hum Genet* **103**, 211-27. (1998).
76. Terwilliger, J. D., Zollner, S., Laan, M. & Paabo, S. **Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion**. *Hum Hered* **48**, 138-54. (1998).
77. Peltonen, L. **Positional cloning of disease genes: advantages of genetic isolates**. *Hum Hered* **50**, 66-75. (2000).
78. Puffenberger, E. G. et al. **Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22**. *Hum Mol Genet* **3**, 1217-25. (1994).
79. Peltonen, L., Palotie, A. & Lange, K. **Use of population isolates for mapping complex traits**. *Nat Rev Genet* **1**, 182-90. (2000).
80. Gulcher, J., Helgason, A. & Stefansson, K. **Genetic homogeneity of Icelanders**. *Nat Genet* **26**, 395. (2000).
81. Wright, A. F., Carothers, A. D. & Pirastu, M. **Population choice in mapping genes for complex diseases**. *Nat Genet* **23**, 397-404. (1999).
82. Jorde, L. B., Watkins, W. S., Kere, J., Nyman, D. & Eriksson, A. W. **Gene mapping in isolated populations: new roles for old friends?** *Hum Hered* **50**, 57-65. (2000).
83. Shifman, S. & Darvasi, A. **The value of isolated populations**. *Nat Genet* **28**, 309-10. (2001).
84. Eaves, I. A. et al. **The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes**. *Nat Genet* **25**, 320-3. (2000).
85. Taillon-Miller, P. et al. **Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28**. *Nat Genet* **25**, 324-8. (2000).
86. Daly, A. K. & Day, C. P. **Candidate gene case-control association studies: advantages and potential pitfalls**. *Br J Clin Pharmacol* **52**, 489-99. (2001).
87. Witte, J. S., Elston, R. C. & Cardon, L. R. **On the relative sample size required for multiple comparisons**. *Stat Med* **19**, 369-72. (2000).

88. Cardon, L. R., Idury, R. M., Harris, T. J., Witte, J. S. & Elston, R. C. **Testing drug response in the presence of genetic information: sampling issues for clinical trials**. *Pharmacogenetics* **10**, 503-10. (2000).
89. Schork, N. J. **Power calculations for genetic association studies using estimated probability distributions**. *Am J Hum Genet* **70**, 1480-9. (2002).
90. Hegele, R. A. **SNP judgments and freedom of association**. *Arterioscler Thromb Vasc Biol* **22**, 1058-61. (2002).
91. Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. **Parametric and nonparametric linkage analysis: a unified multipoint approach**. *Am J Hum Genet* **58**, 1347-63. (1996).
92. Lander, E. & Kruglyak, L. **Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results**. *Nat Genet* **11**, 241-7. (1995).
93. Devlin, B. & Roeder, K. **Genomic control for association studies**. *Biometrics* **55**, 997-1004. (1999).
94. Lee, W. C. **Testing for candidate gene linkage disequilibrium using a dense array of single nucleotide polymorphisms in case-parents studies**. *Epidemiology* **13**, 545-51. (2002).
95. Zhao, J. H., Curtis, D. & Sham, P. C. **Model-free analysis and permutation tests for allelic associations**. *Hum Hered* **50**, 133-9. (2000).
96. Theuns, J. et al. **Genetic variability in the regulatory region of presenilin 1 associated with risk for Alzheimer's disease and variable expression**. *Hum Mol Genet* **9**, 325-31. (2000).
97. Selkoe, D. J. **Alzheimer's disease: genes, proteins, and therapy**. *Physiol Rev* **81**, 741-66. (2001).
98. Alzheimer, A. *Allgemeine Zeitschrift für Psychiatrie* (1907).
99. Leon, M. d. *An atlas of Alzheimer's disease* (The Parthenon Publishing Group Inc., 1999).
100. Petersen, R. C. et al. **Mild cognitive impairment: clinical characterization and outcome**. *Arch Neurol* **56**, 303-8. (1999).
101. Keene, J., Hope, T., Fairburn, C. G. & Jacoby, R. **Death and dementia**. *Int J Geriatr Psychiatry* **16**, 969-74. (2001).
102. Dickson, D. W. **The pathogenesis of senile plaques**. *J Neuropathol Exp Neurol* **56**, 321-39. (1997).
103. Sisodia, S. S. & St George-Hyslop, P. H. **gamma-Secretase, Notch, Abeta and Alzheimer's disease: where do the presenilins fit in?** *Nat Rev Neurosci* **3**, 281-90. (2002).
104. Wood, J. G., Mirra, S. S., Pollock, N. J. & Binder, L. I. **Neurofibrillary tangles of Alzheimer disease share antigenic determinants with the axonal microtubule-associated protein tau (tau)**. *Proc Natl Acad Sci U S A* **83**, 4040-3. (1986).
105. Grundke-Iqbal, I. et al. **Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology**. *Proc Natl Acad Sci U S A* **83**, 4913-7. (1986).
106. Olson, M. I. & Shaw, C. M. **Presenile dementia and Alzheimer's disease in mongolism**. *Brain* **92**, 147-56. (1969).
107. Tanzi, R. et al. **Genetic heterogeneity of gene defects responsible for familial Alzheimer disease**. *Genetica* **91**, 255-63. (1993).
108. Schellenberg, G. D. et al. **Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14**. *Science* **258**, 668-71. (1992).
109. Sherrington, R. et al. **Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease**. *Nature* **375**, 754-60. (1995).

110. Wisniewski, T., Golabek, A., Matsubara, E., Ghiso, J. & Frangione, B. **Apolipoprotein E: binding to soluble Alzheimer's beta-amyloid**. *Biochem Biophys Res Commun* **192**, 359-65. (1993).
111. Pericak-Vance, M. A. et al. **Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage**. *Am J Hum Genet* **48**, 1034-50. (1991).
112. van Duijn, C. M. et al. **A population-based study of familial Alzheimer disease: linkage to chromosomes 14, 19, and 21**. *Am J Hum Genet* **55**, 714-27. (1994).
113. Weisgraber, K. H., Rall, S. C., Jr. & Mahley, R. W. **Human E apoprotein heterogeneity. Cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms**. *J Biol Chem* **256**, 9077-83. (1981).
114. Hardy, J. & Selkoe, D. J. **The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics**. *Science* **297**, 353-6. (2002).
115. Schenk, D. et al. **Immunization with amyloid-beta attenuates Alzheimer-disease-like pathology in the PDAPP mouse**. *Nature* **400**, 173-7. (1999).
116. Dodart, J. C. et al. **Immunization reverses memory deficits without reducing brain Abeta burden in Alzheimer's disease model**. *Nat Neurosci* **5**, 452-7. (2002).
117. Landegren, U., Nilsson, M. & Kwok, P. Y. **Reading bits of genetic information: methods for single-nucleotide polymorphism analysis**. *Genome Res* **8**, 769-76. (1998).
118. Brennan, M. D. **High throughput genotyping technologies for pharmacogenomics**. *Am J Pharmacogenomics* **1**, 295-302 (2001).
119. Breen, G. **Novel and alternate SNP and genetic technologies**. *Psychiatr Genet* **12**, 83-8. (2002).
120. Gut, I. G. **Automation in genotyping of single nucleotide polymorphisms**. *Hum Mutat* **17**, 475-92. (2001).
121. Kwok, P. Y. **High-throughput genotyping assay approaches**. *Pharmacogenomics* **1**, 95-100. (2000).
122. Kwok, P. Y. **Methods for genotyping single nucleotide polymorphisms**. *Annu Rev Genomics Hum Genet* **2**, 235-58 (2001).
123. Shi, M. M. **Technologies for individual genotyping: detection of genetic polymorphisms in drug targets and disease genes**. *Am J Pharmacogenomics* **2**, 197-205 (2002).
124. Syvanen, A. C. **Accessing genetic variation: genotyping single nucleotide polymorphisms**. *Nat Rev Genet* **2**, 930-42. (2001).
125. Tsuchihashi, Z. & Dracopoli, N. C. **Progress in high throughput SNP genotyping methods**. *Pharmacogenomics J* **2**, 103-10 (2002).
126. Warrington, J. A. et al. **New developments in high-throughput resequencing and variation detection using high density microarrays**. *Hum Mutat* **19**, 402-9. (2002).
127. Orita, M., Suzuki, Y., Sekiya, T. & Hayashi, K. **Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction**. *Genomics* **5**, 874-9. (1989).
128. Myers, R. M., Lumelsky, N., Lerman, L. S. & Maniatis, T. **Detection of single base substitutions in total genomic DNA**. *Nature* **313**, 495-8. (1985).
129. Mullis, K. B. & Faloona, F. A. **Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction**. *Methods Enzymol* **155**, 335-50 (1987).
130. Prince, J. A. et al. **Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation**. *Genome Res* **11**, 152-62. (2001).
131. Conner, B. J. et al. **Detection of sickle cell beta S-globin allele by hybridization with synthetic oligonucleotides**. *Proc Natl Acad Sci U S A* **80**, 278-82. (1983).

132. Little, D. P., Braun, A., Darnhofer-Demar, B. & Koster, H. **Identification of apolipoprotein E polymorphisms using temperature cycled primer oligo base extension and mass spectrometry**. *Eur J Clin Chem Clin Biochem* **35**, 545-8. (1997).
133. Haff, L. A. & Smirnov, I. P. **Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry**. *Genome Res* **7**, 378-88. (1997).
134. Kwok, P. Y. **Genotyping by mass spectrometry takes flight**. *Nat Biotechnol* **16**, 1314-5. (1998).
135. Li, J. et al. **Single nucleotide polymorphism determination using primer extension and time-of-flight mass spectrometry**. *Electrophoresis* **20**, 1258-65. [pii] (1999).
136. Sauer, S. et al. **A novel procedure for efficient genotyping of single nucleotide polymorphisms**. *Nucleic Acids Res* **28**, E13. (2000).
137. Sauer, S. et al. **Full flexibility genotyping of single nucleotide polymorphisms by the GOOD assay**. *Nucleic Acids Res* **28**, E100. (2000).
138. Buetow, K. H. et al. **High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip- based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry**. *Proc Natl Acad Sci U S A* **98**, 581-4. (2001).
139. Steen, H. & Mann, M. **Analysis of bromotryptophan and hydroxyproline modifications by high- resolution, high-accuracy precursor ion scanning utilizing fragment ions with mass-deficient mass tags**. *Anal Chem* **74**, 6230-6. (2002).
140. Haff, L. A. & Smirnov, I. P. **Multiplex genotyping of PCR products with MassTag-labeled primers**. *Nucleic Acids Res* **25**, 3749-50. (1997).
141. Park, S. J., Taton, T. A. & Mirkin, C. A. **Array-based electrical detection of DNA with nanoparticle probes**. *Science* **295**, 1503-6. (2002).
142. Galvin, P. **A nanobiotechnology roadmap for high-throughput single nucleotide polymorphism analysis**. *Psychiatr Genet* **12**, 75-82. (2002).
143. Ronaghi, M., Uhlen, M. & Nyren, P. **A sequencing method based on real-time pyrophosphate**. *Science* **281**, 363, 365. (1998).
144. Nyren, P., Karamohamed, S. & Ronaghi, M. **Detection of single-base changes using a bioluminometric primer extension assay**. *Anal Biochem* **244**, 367-73. (1997).
145. Ronaghi, M. **Pyrosequencing for SNP genotyping**. *Methods Mol Biol* **212**, 189-95 (2003).
146. Kwok, P. Y. **SNP genotyping with fluorescence polarization detection**. *Hum Mutat* **19**, 315-23. (2002).
147. Livak, K. J., Flood, S. J., Marmaro, J., Giusti, W. & Deetz, K. **Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization**. *PCR Methods Appl* **4**, 357-62. (1995).
148. Livak, K. J. **Allelic discrimination using fluorogenic probes and the 5' nuclease assay**. *Genet Anal* **14**, 143-9. (1999).
149. Tyagi, S. & Kramer, F. R. **Molecular beacons: probes that fluoresce upon hybridization**. *Nat Biotechnol* **14**, 303-8. (1996).
150. Tyagi, S., Bratu, D. P. & Kramer, F. R. **Multicolor molecular beacons for allele discrimination**. *Nat Biotechnol* **16**, 49-53. (1998).
151. Marras, S. A., Kramer, F. R. & Tyagi, S. **Genotyping SNPs with molecular beacons**. *Methods Mol Biol* **212**, 111-28. (2003).

152. Marras, S. A., Kramer, F. R. & Tyagi, S. **Multiplex detection of single-nucleotide variations using molecular beacons**. *Genet Anal* **14**, 151-6. (1999).
153. Solinas, A. et al. **Duplex Scorpion primers in SNP analysis and FRET applications**. *Nucleic Acids Res* **29**, E96. (2001).
154. Whitcombe, D., Theaker, J., Guy, S. P., Brown, T. & Little, S. **Detection of PCR products using self-probing amplicons and fluorescence**. *Nat Biotechnol* **17**, 804-7. (1999).
155. Thelwell, N., Millington, S., Solinas, A., Booth, J. & Brown, T. **Mode of action and application of Scorpion primers to mutation detection**. *Nucleic Acids Res* **28**, 3752-61. (2000).
156. Pease, A. C. et al. **Light-generated oligonucleotide arrays for rapid DNA sequence analysis**. *Proc Natl Acad Sci U S A* **91**, 5022-6. (1994).
157. Heller, M. J., Forster, A. H. & Tu, E. **Active microeletronic chip devices which utilize controlled electrophoretic fields for multiplex DNA hybridization and other genomic applications**. *Electrophoresis* **21**, 157-64. (2000).
158. Gilles, P. N., Wu, D. J., Foster, C. B., Dillon, P. J. & Chanock, S. J. **Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips**. *Nat Biotechnol* **17**, 365-70. (1999).
159. Shumaker, J. M., Metspalu, A. & Caskey, C. T. **Mutation detection by solid phase primer extension**. *Hum Mutat* **7**, 346-54 (1996).
160. Khanna, M. et al. **Multiplex PCR/LDR for detection of K-ras mutations in primary colon tumors**. *Oncogene* **18**, 27-38. (1999).
161. Iannone, M. A. et al. **Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry**. *Cytometry* **39**, 131-40. (2000).
162. Cai, H. et al. **Flow cytometry-based minisequencing: a new platform for high-throughput single-nucleotide polymorphism scoring**. *Genomics* **66**, 135-43. (2000).
163. Chen, J. et al. **A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension**. *Genome Res* **10**, 549-57. (2000).
164. Healey, B. G., Matson, R. S. & Walt, D. R. **Fiberoptic DNA sensor array capable of detecting point mutations**. *Anal Biochem* **251**, 270-9. (1997).
165. Sauer, S. et al. **Facile method for automated genotyping of single nucleotide polymorphisms by mass spectrometry**. *Nucleic Acids Res* **30**, e22. (2002).
166. Lyamichev, V. et al. **Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes**. *Nat Biotechnol* **17**, 292-6. (1999).
167. Hall, J. G. et al. **Sensitive detection of DNA polymorphisms by the serial invasive signal amplification reaction**. *Proc Natl Acad Sci U S A* **97**, 8272-7. (2000).
168. Griffin, T. J., Hall, J. G., Prudent, J. R. & Smith, L. M. **Direct genetic analysis by matrix-assisted laser desorption/ionization mass spectrometry**. *Proc Natl Acad Sci U S A* **96**, 6301-6. (1999).
169. Hsu, T. M., Law, S. M., Duan, S., Neri, B. P. & Kwok, P. Y. **Genotyping single-nucleotide polymorphisms by the invader assay with dual-color fluorescence polarization detection**. *Clin Chem* **47**, 1373-7. (2001).
170. Nikiforov, T. T. et al. **Genetic Bit Analysis: a solid phase method for typing single nucleotide polymorphisms**. *Nucleic Acids Res* **22**, 4167-75. (1994).
171. Pastinen, T., Kurg, A., Metspalu, A., Peltonen, L. & Syvanen, A. C. **Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays**. *Genome Res* **7**, 606-14. (1997).

172. Shuber, A. P. et al. **High throughput parallel analysis of hundreds of patient samples for more than 100 mutations in multiple disease genes**. *Hum Mol Genet* **6**, 337-47. (1997).
173. Day, I. N., Humphries, S. E., Richards, S., Norton, D. & Reid, M. **High-throughput genotyping using horizontal polyacrylamide gels with wells arranged for microplate array diagonal gel electrophoresis (MADGE)**. *Biotechniques* **19**, 830-5. (1995).
174. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. **A ligase-mediated gene detection technique**. *Science* **241**, 1077-80. (1988).
175. Baron, H. et al. **Oligonucleotide ligation assay (OLA) for the diagnosis of familial hypercholesterolemia**. *Nat Biotechnol* **14**, 1279-82. (1996).
176. Nickerson, D. A. et al. **Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay**. *Proc Natl Acad Sci U S A* **87**, 8923-7. (1990).
177. Grossman, P. D. et al. **High-density multiplex detection of nucleic acid sequences: oligonucleotide ligation assay and sequence-coded separation**. *Nucleic Acids Res* **22**, 4527-34. (1994).
178. Antson, D. O., Isaksson, A., Landegren, U. & Nilsson, M. **PCR-generated padlock probes detect single nucleotide variation in genomic DNA**. *Nucleic Acids Res* **28**, E58. (2000).
179. Nilsson, M. et al. **Making ends meet in genetic analysis using padlock probes**. *Hum Mutat* **19**, 410-5. (2002).
180. Lizardi, P. M. et al. **Mutation detection and single-molecule counting using isothermal rolling-circle amplification**. *Nat Genet* **19**, 225-32. (1998).
181. Brookes, A. J., Howell, W. M., Woodburn, K., Johnstone, E. C. & Carothers, A. **Presenilin-I, presenilin-II, and VLDL-R associations in early onset Alzheimer's disease**. *Lancet* **350**, 336-7. (1997).
182. St Clair, D. et al. **Apolipoprotein E epsilon 4 allele is a risk factor for familial and sporadic presenile Alzheimer's disease in both homozygote and heterozygote carriers**. *J Med Genet* **32**, 642-4. (1995).
183. Woodburn, K. & Johnstone, E. **Ascertainment of a population of people with early-onset dementia in Lothian, Scotland**. *Int J Geriatr Psychiatry* **14**, 362-7. (1999).
184. Strittmatter, W. J. et al. **Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease**. *Proc Natl Acad Sci U S A* **90**, 1977-81. (1993).
185. Wragg, M., Hutton, M. & Talbot, C. **Genetic association between intronic polymorphism in presenilin-1 gene and late-onset Alzheimer's disease. Alzheimer's Disease Collaborative Group**. *Lancet* **347**, 509-12. (1996).
186. Okuizumi, K. et al. **Genetic association of the very low density lipoprotein (VLDL) receptor gene with sporadic Alzheimer's disease**. *Nat Genet* **11**, 207-9. (1995).
187. Emahazion, T. et al. **Identification of 167 polymorphisms in 88 genes from candidate neurodegeneration pathways**. *Gene* **238**, 315-24. (1999).
188. Ylitalo, N., Bergstrom, T. & Gyllensten, U. **Detection of genital human papillomavirus by single-tube nested PCR and type-specific oligonucleotide hybridization**. *J Clin Microbiol* **33**, 1822-8. (1995).
189. Pericak-Vance, M. A. et al. **Identification of novel genes in late-onset Alzheimer's disease**. *Exp Gerontol* **35**, 1343-52. (2000).
190. Howell, W. M., Jobs, M., Gyllensten, U. & Brookes, A. J. **Dynamic allele-specific hybridization. A new method for scoring single nucleotide polymorphisms**. *Nat Biotechnol* **17**, 87-8. (1999).

191. Jobs, M., Howell, W. M. & Brookes, A. J. **Creating arrays by centrifugation**. *Biotechniques* **32**, 1322-4, 1326, 1329. (2002).
192. Nilsson, M. et al. **Padlock probes:circularizing oligonucleotides for localized DNA detection**. *Science* **265**, 2085-8. (1994).

Website references

dbSNP	www.ncbi.nlm.nih.gov/SNP/
HGVbase	hgvbase.cgb.ki.se
OMIM	www.ncbi.nlm.nih.gov/omim/
Retinal information network	www.sph.uth.tmc.edu/Retnet/disease.htm
Cystic fibrosis mutation database	www.sickkids.on.ca/cftr/
Online statistics tools for geneticists	members.aol.com/johnp71/javastat.html#CrossTabs
ALFRED	info.med.yale.edu/genetics/kkidd/