Genetic Polymorphism and SNPs
Genotyping, Haplotype Assembly Problem, Haplotype Map, Functional Genomics and Proteomics
February 19, 2002
Prepared by Kaleigh Smith

1 Introduction to SNPs

The following document provides an introduction single nucleotide polymorphisms and the motivation they provide for current research in pharmacogenomics (and other areas) and the technology to facilitate this research. The main initiative behind SNP-related work is that genetic differences between people to be used to predict phenotypes and phylogeny. We will generally discus SNPs as they occur in human genomic sequences unless specified otherwise. This text is accompanied by a set of slides on SNPs also available at this location.

1.1 Genetic Polymorphism

Genetic Polymorphism A difference in DNA sequence among individuals, groups, or populations. Sources include SNPs, sequence repeats, insertions, deletions and recombination. (e.g. a genetic polymorphism might give rise to blue eyes versus brown eyes, or straight hair versus curly hair). Genetic polymorphisms may be the result of chance processes, or may have been induced by external agents (such as viruses or radiation). If a difference in DNA sequence among individuals has been shown to be associated with disease, it will usually be called a genetic mutation. Changes in DNA sequence which have been confirmed to be caused by external agents are also generally called "mutations" rather than "polymorphisms."[Source: PHRMA Genomics Lexicon]

Genetic Mutation A change in the nucleotide sequence of a DNA molecule. Genetic mutations are a kind of genetic polymorphism. The term "mutation," as opposed to "polymorphism," is generally used to refer to changes in DNA sequence which are not present in most individuals of a species and either have been associated with disease (or risk of disease) or have resulted from damage inflicted by external agents (such as viruses or radiation). [Source: PHRMA Genomics Lexicon]

1.2 SNPs

A Single Nucleotide Polymorphism is a source variance in a genome. A SNP ("snip") is a single base mutation in DNA. SNPs are the most simple form and most common source of genetic polymorphism in the human genome (90% of all human DNA polymorphisms).

There are two types of nucleotide base substitutions resulting in SNPs:

- A **transition** substitution occurs between purines (A, G) or between pyrimidines (C, T). This type of substitution constitutes two thirds of all SNPs.

- A **transversion** substitution occurs between a purine and a pyrimidine.

Sequence Variation
Sequence variation caused by SNPs can be measured in terms of nucleotide diversity, the ratio of

the number of base differences between two genomes over the number of bases compared. This is approximately 1/1000 (1/1350) base pairs between two equivalent chromosomes.

Distribution of SNPs
SNPs are not uniformly distributed over the entire human genome, neither over all chromosomes and neither within a single chromosome. There are one third as many SNPs within coding regions as non-coding region SNPs. It has also been shown that sequence variation is much lower for the sex chromosomes. Within a single chromosome, SNPs can be concentrated about a specific region, usually implying a region of medical or research interest. For instance, the sequence that encodes proteins that present antigens to the immune system in chromosome 6 displays very high nucleotide diversity compared to the other areas of that chromosome.

1.3 Coding Region SNPs

A SNP in a coding region may have two different effects on the resulting protein:

Synonymous the substitution causes no amino acid change to the protein it produces. This is also called a silent mutation.

Non-Synonymous the substitution results in an alteration of the encoded amino acid. A missense mutation changes the protein by causing a change of codon. A nonsense mutation results in a misplaced termination codon. One half of all coding sequence SNPs result in non-synonymous codon changes.

SNPs may also occur in regulatory regions of genes. These SNPs are capable of changing the amount or timing of a protein's production. Such SNPs are much more difficult to find and understand and gene regulation itself is not yet clearly understood.

1.4 Phenotype, Genotype and Haplotype

Phenotype, genotype and haplotype are the most important and basic concepts related to SNPs. It is important to have a clear understanding of each term and the processes of genotyping and haplotyping.

Phenotype The observable properties of an individual as they have developed under the combined influences of the individual's genotype and the effects of environmental factors. [Source: Purves et al. Life: The Science of Biology]

Genotype An exact description of the genetic constitution of an individual, with respect to a single trait or a larger set of traits. [Source: Purves et al. Life: The Science of Biology]. The genetic constitution of an organism as revealed by genetic or molecular analysis, i.e. the complete set of genes, both dominant and recessive, possessed by a particular cell or organism. [IUPAC Biotech]

Genotyping Genotyping is normally defined as detecting the genotypes of individual SNPs. In diploid organisms (have alternative alleles of SNPs), such as humans, the linkage of particular SNP genotypes on each chromosome in a homologous pair (the haplotype) may provide additional information not available from SNP genotyping alone. [CHI SNP Update]

Haplotype (haploid genotype) A particular pattern of sequential SNPs (or alleles) found on a single chromosome. These SNPs tend to be inherited together over time and can serve as disease- gene markers. The examination of single chromosome sets (haploid sets), as opposed to the usual chromosome pairings (diploid sets), is important because mutations in one copy of a chromosome pair can be masked by normal sequences present on the other copy [CHI SNP Update]. A combination of alleles of closely linked loci that are found in a single chromosome and tend to be inherited together. The linear, ordered arrangement of alleles on a chromosome. Haplotype analysis is useful in identifying recombination events. [Sources: Purves et al. Life: The Science of Biology and PHRMA Genomics Lexicon]

Haplotyping Haplotyping involves grouping subjects by haplotypes, or particular patterns of sequential SNPs, found on a single chromosome. [CHI SNPs Update]

Genomic variation, and thus SNPs, is responsible for diversity in the human species. It follows that since SNPs account for diversity in human genotypes, they can be mapped to account for diversity in phenotypes. An "individual SNPs may serve as signposts for disease genes, haplotypes are believed to be superior for this purpose The study of haplotypes within genes, which is also of great current interest, provides the opportunity to discover reliable markers of various phenotypes" [CHI SNP Update]. This relation forms the basis and motivation for the identification and genotyping of SNPs.

2 SNP Discovery and SNP Genotyping

2.1 SNP Detection or SNP Discovery

There are over one million SNPs identified (1,255,326 mapped SNPs at the SNP Consortium Organization). Validation experiments have shown that 95% of these are unique and valid polymorphisms (not the product of error or redundancy).

Methods for SNP discovery/detection involve a set of biochemical reactions that isolates the precise location of a suspected SNP and then directly determines the identity of the SNP, using an enzyme called DNA polymerase. [Source: http://www.orchid.com/]

Also, many SNPs were initially detected by comparing different sequenced genomes. This work has now been extended to a much larger-scale effort to determine the SNPs (genotypes) of many genomes from different populations.

Notice the difference between SNP discovery/detection and SNP scoring or SNP genotyping. One strives to identify new SNP locations on the genome, while the other "involves methods to determine the genotypes of many individuals for particular SNPs that have already been discovered" [NIH, Methods for Discovering and Scoring Single Nucleotide Polymorphisms, Request for Applications Jan. 9, 1998]. This ends our discussion of SNP detection. What follows is an overview of "post-genomic" SNP related applications such as high throughput genotyping, determining haplotypes from genotypes, and haplotype mapping.

2.2 High-Throughput SNP Genotyping

The second phase of human genomics (the first being the sequencing of the human genome) involves large-scale screening of different human populations for significant DNA polymorphisms. The in-

formation gathered will lead to accurate associations between genotype and phenotype.

High-throughput SNP genotyping is the process of quickly and cost-effectively identifying the SNP values in as many different individual human genomes as possible. Steps of SNP genotyping involve DNA sample preparation, PCR amplification, and microarray assays. For the last step, the technology must label SNP locations of both alleles in the DNA sample and determine the base values using microarray technology.

Orchid Biocomputer and Affymetrix are the leaders in providing SNP genotyping technology. They have developed single nucleotide polymorphism (SNP) genotyping assays that combine Orchids proprietary GBA primer extension technology with an Affymetrix GeneChip universal array. Their technologies can provide ultra-high throughput of 100,000 genotypes/day. It is interesting to note that nearly all SNP genotyping uses Affymetrix equipment (including GenFlex, Affymetrix's universal microarray).

High-throughput SNP genotyping achieves the goal of recording the SNP location base values of thousands of people from a given population. The genotype data gathered from high-throughput genotyping is then used to determine the related haplotypes. This information is then used for SNP mapping which is discussed further on in this summary.

3   The SNP Haplotype Assembly Problem

SNP genotyping has introduced a complex computational problem (luckily). This problem was first published by Lancia et al. in the paper "SNPs Problems, Complexity and Algorithms", and followed by "Algorithmic Strategies for the single nucleotide polymorphism haplotype assembly problem" by Lipert et al. Before understanding the motivation for this problem, it is important to be familiar with the following terms.

Diploid The chromosome complement consists of two copies (homologues) of each chromosome. In humans, the each chromosome pair is from a different origin (mother, father). [Source: Purves et al. Life: The Science of Biology]

Allele The alternative forms of a genetic character found at a given locus on a chromosome. [Source: Purves et al. Life: The Science of Biology]

Homozygous A diploid organism having identical alleles of a given gene on both homologous chromosomes. [Source: Purves et al. Life: The Science of Biology]

Heterozygous A diploid organism having different alleles of a given gene on both homologous chromosomes. [Source: Purves et al. Life: The Science of Biology]

Haplotype A combination of alleles of closely linked loci that are found in a single chromosome and tend to be inherited together. The linear, ordered arrangement of alleles on a chromosome. [Sources: Purves et al. Life: The Science of Biology and PHRMA Genomics Lexicon]

The SNP haplotype assembly problem involves determining the haplotypes from genomic sequence fragments determined by SNP genotyping. For more information on haplotypes, see the following "Haplotype Map" section. Haplotypes can also be determined directly from genomic DNA.

However, this is sometimes more costly and slow than a computational method for determining haplotypes.

Large-scale human genotyping technology introduces an interesting algorithmic problem of partitioning SNP genotype data into haplotype partitions. The problem arises because the genomic fragments that constitute the genotype of a diploid organism contain two copies of each location or chromosome (two haplotypes). The polymorphic DNA fragments must then be assembled into their original haplotype.

The papers describe algorithms to determine haplotypes over long regions from short sequence fragments. The aligned fragments must be partitioned into two sets according to their original homologue. This partitioning is accomplished by using conflicting SNP values between fragments to create a SNP conflict graph or a fragment conflict graph. The following is a summary of the two aforementioned papers.

The SNP Assembly Problem (Lancia)
A SNP assembly is a tuple (S, F, R) where S is a set of n SNPs, F is a set of m fragments and R is a relation $R : SxF \rightarrow \{0, A, B\}$ indicating whether a SNP $s \in S$ does not occur on a fragment $f \in F$ (marked by 0) or if occurring, the "score" of s (A or B).

Computational Haplotyping (Lancia)
A partition of F into two blocks: H_1 and H_2 called haplotypes. A SNP assembly is 'feasible' when there exists a haplotyping such that:

$$\forall s \in S \text{ and } \forall f, f' \in H_i: \ R(s, f) = R(s, f') \text{ or } R(s, f) = 0 \text{ or } R(s, f') = 0.$$

A present SNP $s \in S$ has value either A or B. Recall that the organism is diploid and there are copies of that SNP in fragments taken from both homologues of the sample's chromosome. If the SNP is heterozygous, then the SNP will have value A in H_1 and B in H_2 (or vice versa). If the SNP is homozygous, it can not be used to help determine the fragment's haplotype and is therefore not considered.
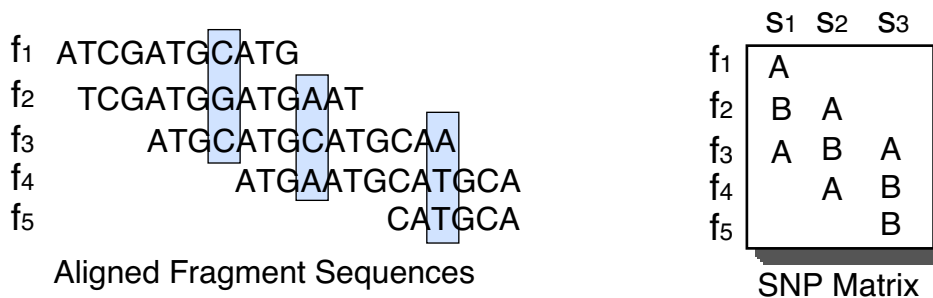

Figure 1: The papers define a SNP matrix as shown.

S and F are both used to create **conflict graphs** G_S and G_F so that an algorithm can determine the fewest $s \in S$ or $f \in F$ that must be removed as errors to render the graphs conflict-free. Two fragments f_i and f_j are in conflict when the exists a SNP $s \in S$ such that $R(f_i, s) \neq 0$ and

$R(f_j, s) \neq 0$ and $R(f_i, s) \neq R(f_j, s)$. Two SNPs s_1 and s_2 are in conflict when there exist two fragments f_1 and f_2 such that three of $R(f_1, s_1)$, $R(f_2, s_1)$, $R(f_1, s_2)$ or $R(f_2, s_2)$ have the same non-zero value and one has the opposing non-zero value.
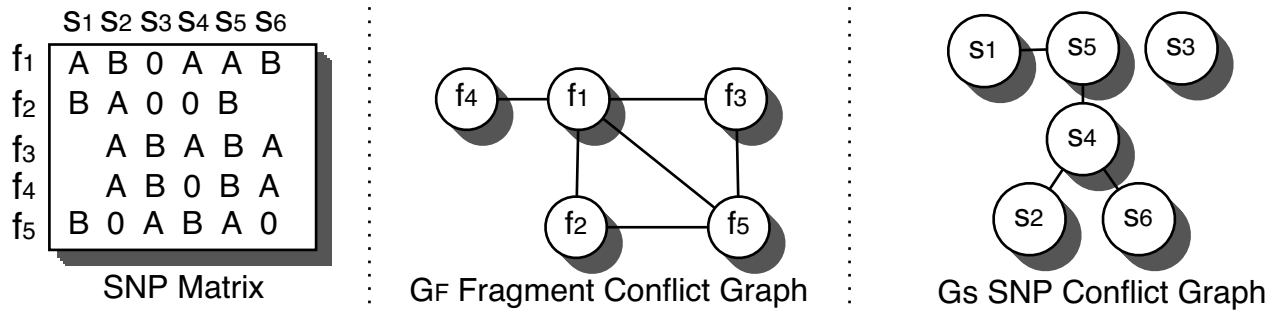

Figure 2: Example SNP matrix and the associated fragment conflict graph and SNP conflict graph.

The SNP haplotype assembly problem has now become either the Minimum Fragment Removal Problem or the Minimum SNP Removal Problem. Both these problems take G_F or G_S respectively as input and return a maximally large G'_F that is bipartite (MAX Induced Bipartite Subgraph problem) or a maximally large G'_S that is a stable set (Vertex Cover). Note that G_F is bipartite iff G_S is a stable set. Both of these problems are NP-hard.
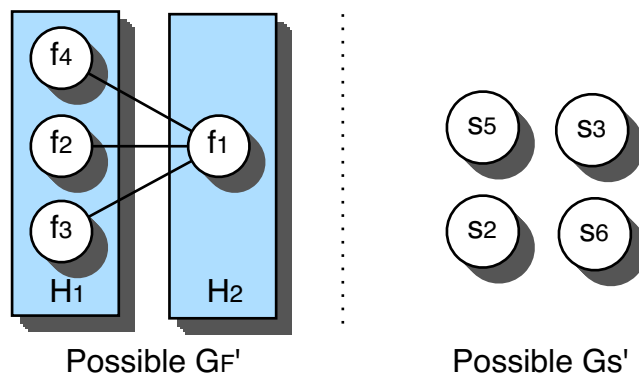

Figure 3: Possible G'_F and G'_S for the above example.

Maximum stable set (Vertex Cover) is a very well studied problem that can be attacked using one of many approximation algorithms. We will therefore focus on the problem of finding the largest bipartite subgraph in G_F. A graph G is bipartite iff it contains no odd cycles. We will therefore consider the MFR problem in terms of a cover problem, the Odd Cycle Cover problem.

Clearly, OCC determines the minimal set of vertices F' to remove from G_F to remove all odd cycles, and therefore render the graph bipartite. OCC has a 9/4-approximation algorithm. It has not yet been shown Fixed Parameter Tractable (FPT). We would like to consider an algorithm
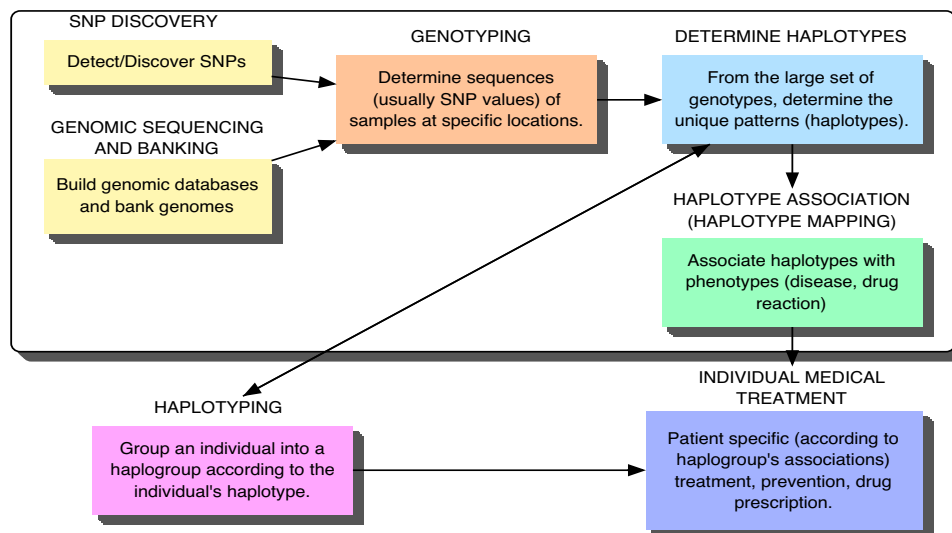
with bounded exponential time to solve a parameterized version of this problem, $k - OCC$. The problem of k-**Odd Cycle Cover** is as follows: Given a graph $G = (V, E)$, find V', $|V'| \leq k$ and $|V'|$ is minimal such that $G' = V - V'$ is bipartite. This is the central function to establish when solving MFR. The second paper describes a naive branch and bound approach based on searching a tree of possible solutions. Our approach involves creating a set of reduction rules to condense $G_F \rightarrow G_F^R$, making it possible to apply efficient (parameterized) algorithms to find a small ($< k$) odd cycle cover of G_F^R, to determine that G_F^R contains greater than k disjoint odd cycles (and therefore does not have a small enough OCC) or contains one other graph structure that disallows a small enough OCC.

This is a work in progress. However, it is good to read Lancia's papers to understand the algorithmic problems related to SNP haplotyping. Let it be known that there are strong statistical methods for performing haplotyping, and as the costs of SNP assays go down, there are also biological methods for determining the haplotypes.

4   SNP Mapping

SNP mapping (association mapping) is one of the most active areas of SNP post-genomics research. This work involves identifying SNP sites along the genome to track disease genes. A human SNP map specifies the contributions of individual genes to diseases and other phenotypes.

The SNP Consortium Ltd. is a non-profit foundation that provides over one million detected SNPs and their annotations to the public. Its mission is to develop up to 300,000 SNPs distributed evenly throughout the human genome and to make the information related to these SNPs available to the public without intellectual property restrictions. http://snp.cshl.org/

Figure 4: The schematic of SNP Mapping (Haplotype Mapping) and one of its many influences on health care.

4.1 A Haplotype Map for the Human Genome

SNP mapping succeeds in identifying individual genes responsible for monogenic diseases such as Huntingtons and cystic fibrosis, however the majority of traits are influenced by multiple genes and environmental factors. As an extension to basic SNP mapping, human genetic variation research determines how variation among individuals or groups contributes to the health status of that individual or group. This type of research has the initiative of developing a **haplotype map** of the human genome. The map's purpose is to relate human genetic variation with disease predisposition, specifically common or complex disorders.

Scientists have discovered that there is a small number of different versions of certain genetic blocks (small number of haplotypes). This means that there is a small number of SNP patterns at each chromosomal position. For instance, for some blocks, only four or five patterns of SNPs were found (four or five different haplotypes) that account for 80%-90% of the entire population. This finding greatly simplifies the search for associations between DNA variations and disease.

Also, as many such haplotype patterns are specific to populations (groups), the map will facilitate the conduct of association studies in selected populations where certain diseases are more or less prevalent.

The following information from CHI puts haplotype mapping in context.

Haplotype map Francis Collins, director of the NHGRI, speaking at BIO 2001 (San Diego CA, US, June 2001) announced plans for a public- private effort to create a human haplotype map. Creators hope this so- called haplotype map will be a tool for pinning down the genes that contribute to the development of complex diseases such as cancer, diabetes, and mental illness. [L. Helmuth "Map of the Human Genome 3.0" Science 293 (5530) :583-5 July 27, 2001] [from CHI Glossary].

Haplotype mapping Haplotype mapping is often carried out as part of a genome scan. In a population isolate, the appearance of a rare Mendelian disease is almost always attributable to a single founder gene or mutation. The disease allele can be identified by searching for a common haplotype signature shared among patients. As the ancestral haplotype signature is passed from generation to generation, it is disrupted by recombination. Partial conservation of the haplotype signature in a patient strongly suggests that the disease locus resides in the conserved region of the haplotype. [L. Peltonen et. al, "USE OF POPULATION ISOLATES FOR MAPPING COMPLEX TRAITS" Nature Reviews Genetics 1: 182-190 (2000)

The construction off a haplotype map raises many ethical concerns. Recall that the human genome project faced opposition when it was initially proposed and does still. The ethical issues surrounding a haplotype map are much more severe and delicate. A haplotype map may include markers that indicate someone's race and ethnicity.

4.2 Haplotype Trees (Haplotype-based Phylogeny)

Haplotype trees provide methods for examining the phylogeny of individuals based on their haplotypes and also provide methods for understanding molecular (genomic) natural selection. They are constructed to understand human evolution, historical timelines and to genetically determine genealogy. These trees can be created for one species or can be created to represent inter-species

haplotypic phylogenies, Recall that there are a small number of haplotypes (unique patterns) for chosen location of interest. A population or haplotype group is a set of highly similar haplotypes. Often the haplotype under consideration is a maternally inherited gene or a set of locations on one of the sex chromosomes. It has been shown that members of a population generally share the same haplotype pattern. These trees are often combined with homogy-based trees to provide a more reliable portrait of geneologies.

Haplotype trees are constructed parsimoniously with unique haplotypes being represented by the nodes of the tree. Haplotypes develop from older ancestral haplotypes. These older haplotypes are believed to be more widespread over the species, and are therefore generally represented by internal nodes, whereas newer haplotypes (more recently emerged patterns) will be represented by leaf nodes. Note that it is integral to select a haplotype that is robust to recombination and mutations.
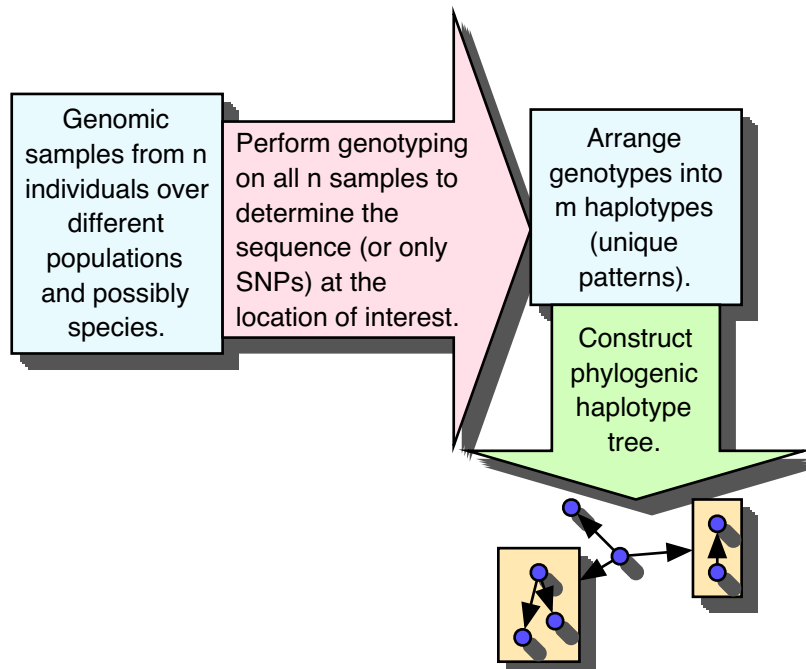

Figure 5: The construction of a haplotype tree

Jin, et al describe the use of a haplotype tree for studying migratory history, genetic drift and natural selection in their 1999 paper "Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations". The haplotypes considered in their research originate from a 565-bp chromosome 21 region near the MX1 gene that contains 12 SNP locations and lack recombination and recurrent mutation events. The paper notes that some haplotypes occur much more frequently than others and that the variety of haplotypes present in different populations vary with older populations displaying higher haplotype variety.

An example of a commercial application of haplotype trees is the Internet company "DNA Family Tree" - http://www.familytreedna.com/, a self-declared "DNA-driven genealogical testing company". The scientists at this company use a set of statistical and phylogenetic (tree-building)
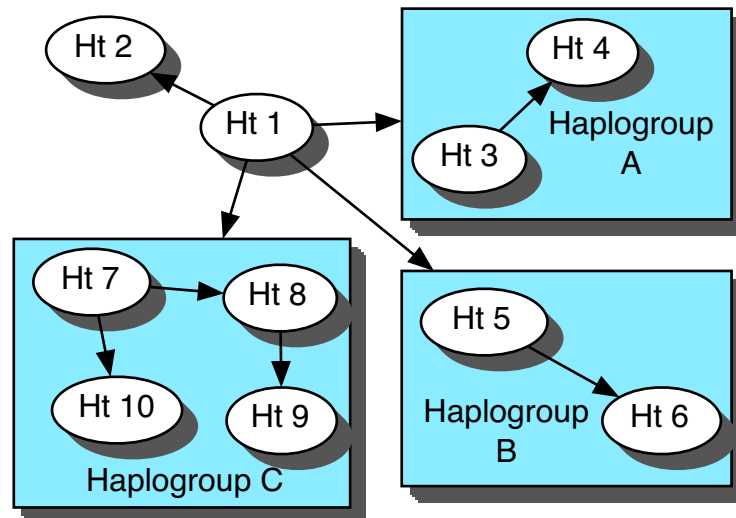
Figure 6: A sample haplotype tree as determined in the above mentioned paper.

methods that show the exact genealogical linkages amongst the haplotypes (either Y chromosome or mitochondrial DNA) to determine the ancient lineage of an interested client. Particularly, one of the products is used to determine if a female client has Native American ancestry. The company can determine this because "genetic studies have shown that Native American mtDNAs belong to one of five distinct maternal lineages. These have been designated haplogroups A, B, C, D and X. Each of these is defined by a specific set of mutations/markers that occur in the coding and non-coding regions of the mtDNA genome." This use of haplotypes again introduces many ethical concerns and it is not difficult to imagine how haplotype technology could facilitate genomic discrimination.

5 SNP Applications

The following are fields that utilize SNP information and haplotype maps. The major application of SNP information is towards improved and futuristic health care. Genomics and specifically SNP research can be used to improve health care through gene therapy, to yield new targets for drug discovery, to refine the process of drug development and to discover new diagnostics.

5.1 Pharmacogenomics

Pharmacogenomics The science of understanding the correlation between an individual patient's genetic make-up (genotype) and their response to drug treatment. Some drugs work well in some patient populations and not as well in others. Studying the genetic basis of patient response to therapeutics allows drug developers to more effectively design therapeutic treatments. [Source: PHRMA Genomics Lexicon]

All aspects of pharmacogenomics require data from high-throughput genotyping, specifically the target population for a drug or the population of people who react poorly with the drug. Also, this type of research may lead to population specific treatments. The high cost of drug recalls have provided an initiative for advanced drug design involving drug-target validation studies as well as

studies to predict adverse events and lack of efficacy.

A sample pharmacogenomic experiment may proceed as follows:

- Define the drug response (phenotype) of interest

- Accumulate patients/DNA/families

- Identify candidate genes that might explain significant response variations

- Identify polymorphisms in candidate genes

- Relate the identified polymorphism to the phenotype

5.2   SNP Diagnostics

Genetic testing The analysis of an individual's genetic material. Among the purposes of genetic testing could be to gather information on an individual's genetic predisposition to particular health condition, or to confirm a diagnosis of genetic disease. [Source: PHRMA Genomics Lexicon]

An individual's genotype can be determined and then analysed according to a haplotype map to determine the patient's disease risk or reception to different treatments.

5.3   SNPs in Functional Proteomics and Gene Therapy

SNP related functional proteomics involve the identification of functional SNPs that modify proteins and protein active sites structure and function. Functional proteomics is closely tied to modern (post-genomic) drug design and function SNP information helps to discover new therapeutic targets. Most interestingly, by developing a database of the modifications generated by functional (coding) SNPs in disease related proteins, "new compounds can be designed for correcting or enhancing the effects of those mutations in the population." [Source: Genodyssee]

What are these compounds and how can knowledge of SNP effects be used to correct populations with undesirable SNPs or enhance populations by introducing the advantages of a desirable SNP? Aside from drugs, here are some interesting genomic therapies that may become more feasible as SNP information in the form of trees and maps become more detailed.

Germ line gene therapy Germ line gene therapy involves the insertion of normal genes into germ cells or fertilized eggs in an attempt to create a beneficial genetic change which can be transmitted to an organism's offspring (for example, to correct for a genetic trait associated with disease). If a change is introduced via germ line gene therapy, that change may be present in the offspring from birth in every cell in the body. See genomics, and compare with somatic cell gene therapy. [Source: PHRMA Genomics Lexicon]

Somatic cell genetic mutation A genetic mutation in a somatic cell. Such mutations, which are not inherited from parents but occur during the lifetime of an organism, are often known as "acquired genetic mutations." Somatic cell genetic mutations are not passed on to offspring. [Source: PHRMA Genomics Lexicon]

Somatic cell gene therapy Somatic cell gene therapy involves the insertion of genes into cells for therapeutic purposes, for example to induce the treated cells to produce a protein that the body is missing. It does not affect the genetic makeup of a patient's offspring, and generally does not change all, or even most, cells in the recipient. [Source: PHRMA Genomics Lexicon]

6   The SNP Economy

The following is an excerpt from the Cambridge Healthtech Institute Article: "SNP-Research Market Could Reach $1.2 Billion By 2005, If Pharmacogenomics Advances January 1, 2002". By Malorye Allison Branca. [Source: http://www.chireports.com/content/articles/snpresearch.asp]

"Annual expenditures on single nucleotide polymorphism (SNP) research could increase seven-fold by 2005, growing to more than $1.2 billion, from $158 million in 2001, according to a new Cambridge Healthtech Institute (CHI) report, Commercial Implications of Advances in the Identification, Mapping, and Application of Single Nucleotide Polymorphisms. (For more on this report, go to www.chireports.com/content/reports/snpupdate01.asp.)"

"Three major factors will fuel this growth:

- Progressive decreases in the cost per genotyping assay, because of technological advances.

- Increasing interest in pharmacogenomics-or tailoring treatment to patients based on their genomic profiles-by pharmaceutical, biotechnology, and genomic tools companies.

- The push to do a greater number of assays per study."

"Currently, the most common applications for SNP-related research tools are gene-disease association studies and drug-target validation. Other popular applications are disease-susceptibility studies or diagnostics, pharmacogenomic studies for clinical trials, drug-target screening, and new technology development. We anticipate that target validation and disease association studies will continue to be the most common SNP-related tasks in drug discovery and development; however, we expect that by 2003, the application of SNP studies in pharmacogenomics will begin to increase steadily and could quickly become a multibillion-dollar market itself."

7   Resources

Affymetrix www.affymetrix.com

Cambridge Healthtech Institute Articles: http://www.chireports.com

Cambridge Healthtech Institute Glossary: http://www.genomicglossaries.com/

DNA Family Tree company: http://www.familytreedna.com/

Lancia, G., Afna, V., Istrail, S., Lippert, R. and Schwartz, R.. *SNPs Problems, Complexity and Algorithms.* ESA 2002, LNCS 2161, pp. 182-193, 2001. Springer-Verlag Berlin Heidelberg 2001.

Jin, Li, et al. *Distribution of haplotypes from a chromosome 21 region distinguishes multiple pre-historic human migrations.* Proc. Natl. Acad. Sci. USA Vol. 96, pp. 3796-3800, March 1999.

Lippert, R., Schwartz, R., Lancia, G. and Istrail, S. *Algorithmic strategies for the single nucleotide polymorphism haplotype assembly problem.* Briefings in Bioinformatics. Volume 3, NO 1, 1-9. February 2002.

Map of the Human Genome 3.0, BioSINO, Laura Helmuth http://www.biosino.org/bioinformatics/010814-4.htm

Orchid BioSciences www.orchid.com

Pharmaceutical Research and Manufacturers of America (PhRMA) Genomics glossary: http://genomics.phrma.

PolyGenyx: http://www.polygenyx.com/

SNP details: SNP Web Source, Xuan Chen (offline)

Studies of the ethical, legal and social implications of human genetic variation research for individuals and diverse racial and ethnic groups: http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-02-003.html