Structural Genomics

Marco F Ramoni, PhD
September 15th, 2005

Biomedical Computing
6.872 / HST 950

Introduction

- On February 12, 2001 the Human Genome Project announces the completion of a first draft of the human genome.
- Among the items on the agenda of the announcement, a statement figures prominently:

A SNP map promises to revolutionize both mapping diseases and tracing human history.

SNP are Single Nucleotide Polymorphisms, subtle variations of the human genome across individuals.

- You can take this sentence as the announcement of a new era for population genetics.

Outline

Properties of the Genome

Basics

- Genetic polymorphisms;
- Evolution and selection;

Genetic diseases

- Tracking genetic diseases;
- Traits and complex traits;

Genomic diseases

- Blocks of heredity;
- Tracking blocks.

The Genetic Study of the Future

Candidates identification

- Find the genes;
- Find the SNPs;

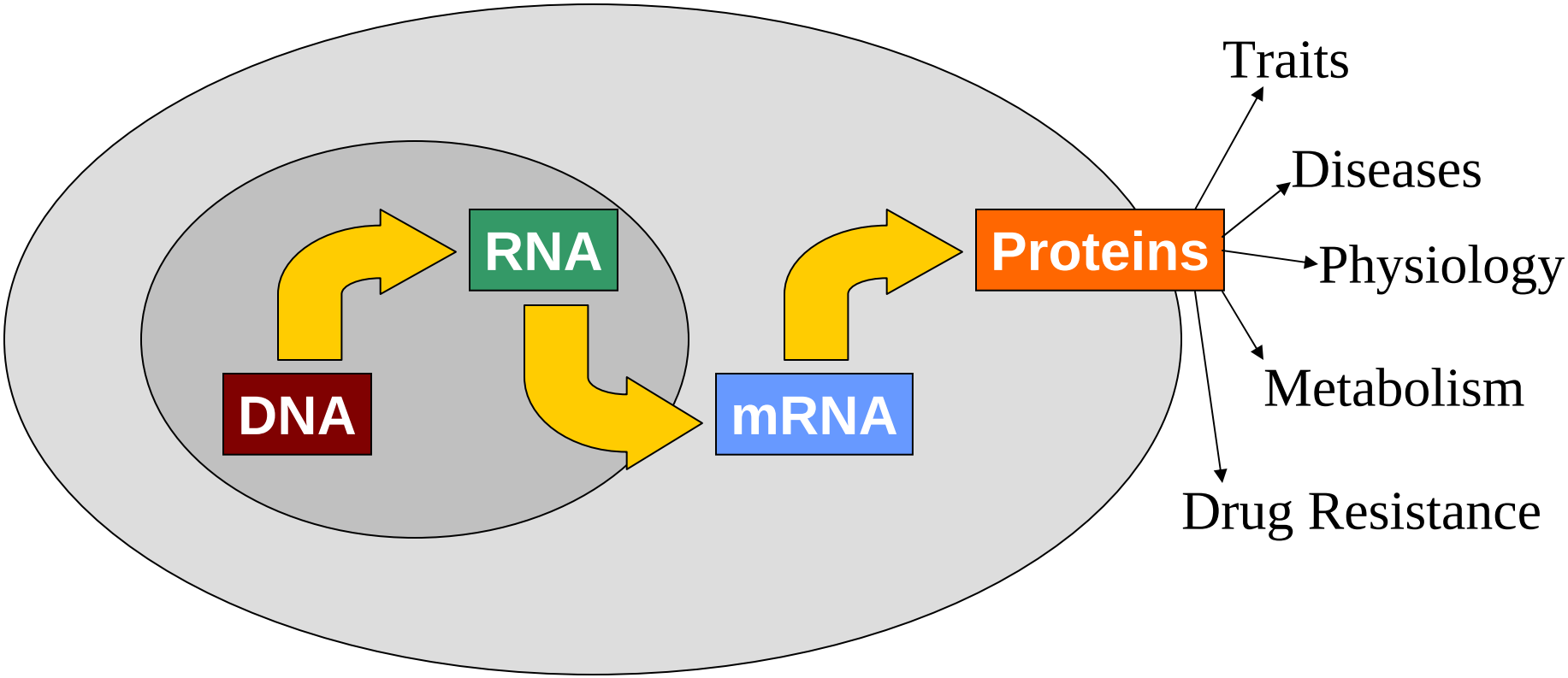
Study design

- Case/control studies;
- Pedigree studies;
- Trios, sibs and TDT;

Study analysis

- Single gene association;
- Multivariate association;
- Validation.

Central Dogma of Molecular Biology



Single Nucleotide Polymorphisms

- Variations of a single base between individuals:

  . . . **ATGCGATCGATACTCGATAACTCCCGA** . . .

  . . . **ATGCGATCGATACGCGATAACTCCCGA** . . .

- A SNP must occur in at least 1% of the population.
- SNPs are the most common type of variations.
- Differently to microsatellites or RTLPs, SNPs may occur in coding regions:
  - cSNP: SNP occurring in a coding region.
  - rSNP: SNP occurring in a regulatory region.
  - sSNP: Coding SNP with no change on amino acid.

Terminology

Allele: A sequence of DNA bases.

Locus: Physical location of an allele on a chromosome.

Linkage: Proximity of two alleles on a chromosome.

Marker: An allele of known position on a chromosome.

Distance: Number of base-pairs between two alleles.

centiMorgan: Probabilistic distance of two alleles.

Phenotype: An outward, observable character (trait).

Genotype: The internally coded, inheritable information.

Penetrance: No. with phenotype / No. with allele.

Distances

- Physical distances between alleles are base-pairs. But the recombination frequency is not constant.

Segregation (Mendel's first law): Allele pairs separate during gamete formation and randomly reform pairs.

- A useful measure of distance is based on the probability of recombination: the Morgan.
- A distance of 1 centiMorgan (cM) between two loci means that they have 1% chances of being separated by recombination.
- A genetic distance of 1 cM is roughly equal to a physical distance of 1 million base pairs (1Mb).

More Terminology

Physical maps: Maps in base-pairs.

Human autosomal physical map: 3000Mb (bases).

Linkage maps: Maps in centiMorgan.

Human Male Map Length: 2851cM.

Human Female Map Length: 4296cM.

Correspondence between maps:

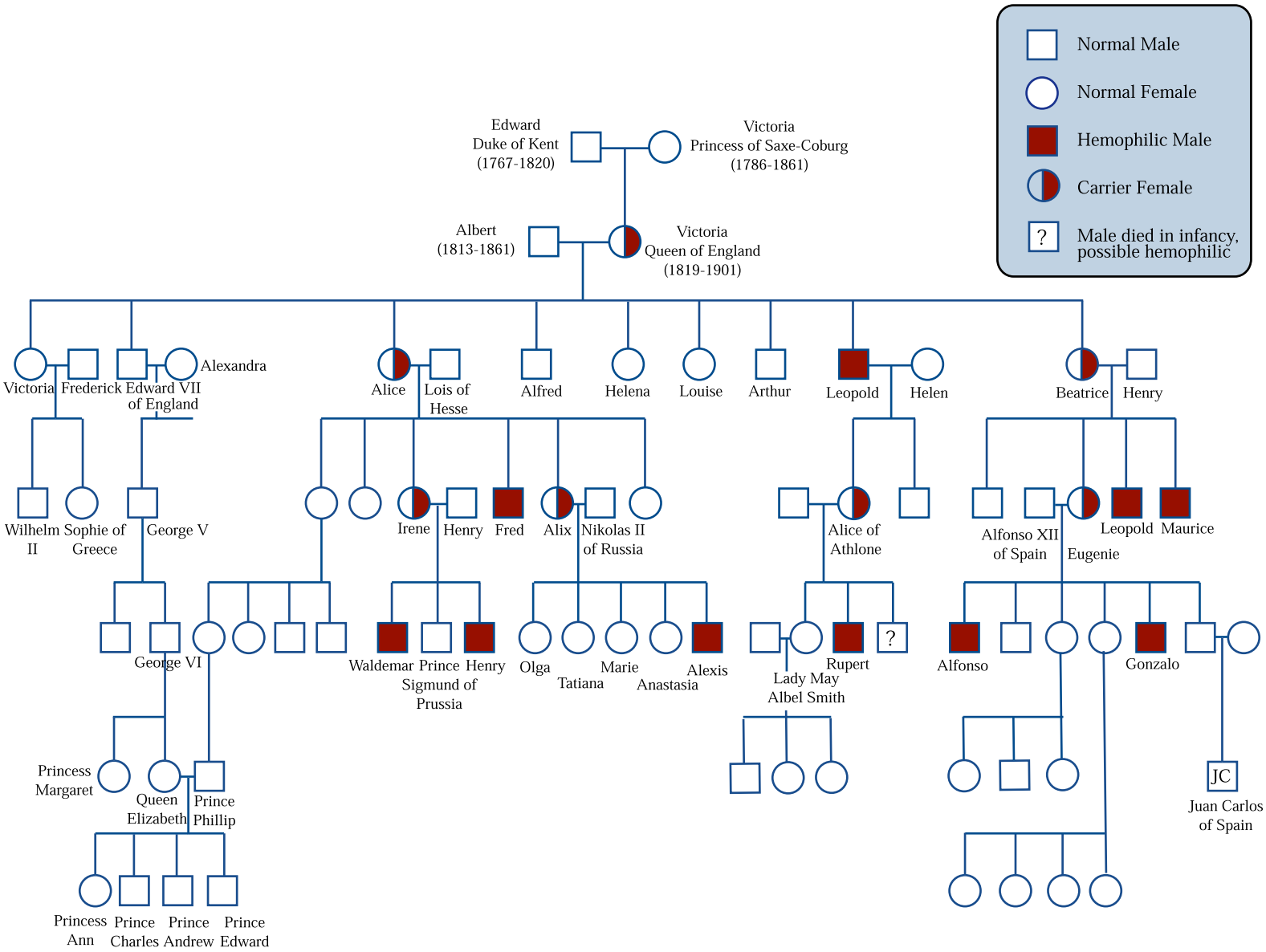
Male cM ~ 1.05 Mb; Female cM ~ 0.88Mb.

Cosegregation: Alleles (or traits) transmitted together.

Hemophilia, a Sex Linked Recessive

- Hemophilia is a X-linked recessive disease, that is fatal for women.
- X-linked means that the allele (DNA code which carries the disease) is on the X-chromosome.
- A woman (XX) can be carrier or non-carrier: if x=allele with disease, then xX=carrier; xx=dies; XX=non carrier.
- A male (YX) can be affected or not affected: (xY= affected; XY=not affected).

Hemophilia: A Royal Disease



Genetic Markers

- One of the most celebrated findings of the human genome project is that humans share most DNA.
- Still, there are subtle variations:

  Simple Sequence Repeats (SSR): Stretches of 1 to 6 nucleotide repeated in tandem.

  Microsatellite**:** Short tandem repeat (e.g. GATA) varying in number between individuals.

  Single nucleotide polymorphism (SNP): Single base variation with at least 1% incidence in population.

Evolutionary Pressure

- Kreitman (1983) sequenced the first 11 alleles from nature: alcohol dehydrogenase locus in Drosophila.
- 11 coding regions / 14 sites have alternative bases.
- 13 variations are silent: ie do not change amino acid.
- With a random base change, we have 75% chances of changing the amino acid (i.e. creating a cSNP).
- Why this disparity?
- Drosophilae and larvae are found in fermenting fruits.
- Alcohol dehydrogenase is important in detoxification.
- A radical change in protein is a killer.

Hardy-Weinberg Law

Hardy-Weinberg Law (1908): Dictates the proportion of major (p), minor alleles (q) in equilibrium.

$$p^2 + 2pq + q^2 = 1.$$

Equilibrium: Hermaphroditic population gets equilibrium in one generation, a sexual population in two.

Example: How many Caucasian carriers of C. fibrosis?

- Affected Caucasians (q^2) = 1/2,500.
- Affected Alleles (q)= 1/50 = 0.02.
- Non Affected Alleles (p) = (1 - 0.02) = 0.98.
- Heterozygous (2pq) = 2(0.98 × 0.02) = 0.04= 1/25.

Assumptions

Random mating: Mating independent of allele.

- Inbreeding: Mating within pedigree;
- Associative mating: Selective of alleles (humans).

Infinite population: Sensible with 6 billions people.

Drift: Allele distributions depend on individuals offspring.

- Locality: Individuals mate locally;
- Small populations: Variations vanish or reach 100%.

Mutations contrast drift by introducing variations.

Heresy: This picture of evolution as equilibrium between drift and mutation does not include selection!

Natural Selection

Example: p=0.6 and q=0.4.

AA	Aa	aa
36%	48%	16%

Fitness (w): AA=Aa=1, aa=0.8. Selection: s = 1-w = 0.2:

$$\delta p = \frac{spq^2}{1-sq^2} = \frac{(0.2)(0.6)(0.4)^2}{1-(0.2)(0.4)^2} = \frac{0.019}{0.968} = 0.02$$

Selection: Effect on the 1st generation is A=0.62 a=0.38.

AA	Aa	aa
39.7%	46.6%	13.7%
+3.7%	-1.4%	-2.3%

Rate: The rate decreases. Variations do not go away.

Does it work?

Race and Sanger (1975) 1279 subjects' blood group.

p = p(M) = (2 x 363) + 634 / (2 x 1279) = 0.53167.

	MM	MN	NN
Observed	363	634	282
Expected	361.54	636.93	280.53

Caveat: Beta-hemoglobin sickle-cell in West Africa:

	AA	AS	SS
Observed	25,374	5,482	64
Expected	25,561.98	5,106.03	254.98

Does it work?

Race and Sanger (1975) 1279 subjects' blood group.

p = p(M) = (2 x 363) + 634 / (2 x 1279) = 0.53167.

	MM	MN	NN
Observed	363	634	282
Expected	361.54	636.93	280.53

Caveat: Beta-hemoglobin sickle-cell in West Africa:

	AA	AS	SS
Observed	25,374	5,482	64
Expected	25,561.98	5,106.03	254.98

Reason: Heterozygous selective advantage: Malaria.

Linkage Equilibrium/Disequilibrium

Linkage equilibrium: Loci Aa and Bb are in equilibrium if transmission probabilities π_A and π_B are independent.

$$\pi_{AB} = \pi_A \pi_B.$$

Haplotype: A combination of allele loci: $\pi_{AB}, \pi_{Ab}, \pi_{aB}, \pi_{ab}$.

Linkage disequilibrium: Loci linked in transmission as.

$$r^2 = \frac{(\pi_{AB} - \pi_A \pi_B)^2}{\pi_A \pi_B \pi_a \pi_b}$$

a measure of dependency between the two loci.

Markers: Linkage disequilibrium is the key of markers.

Phenotype and Genotype

Task: Find basis (genotype) of diseases (phenotype).

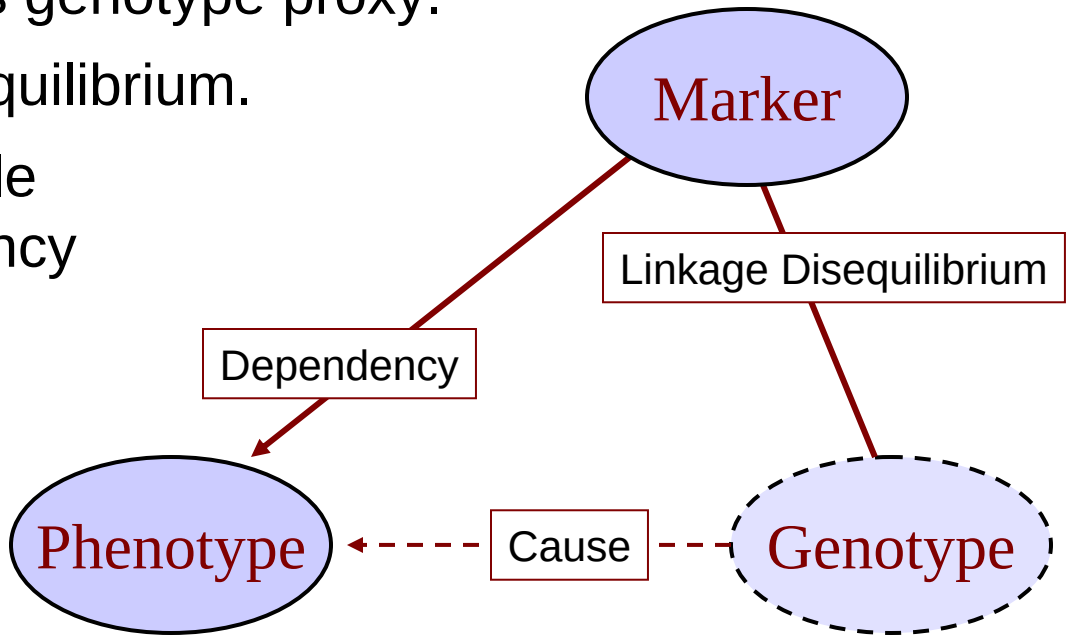
Marker: Flag genomic regions in linkage disequilibrium.

Problem: *Real* genotype is not observable.

Strategy: Use marker as genotype proxy.

Condition: Linkage disequilibrium.

Dependency: Observable measure of dependency between marker and phenotype.



Feasibility: Time and Cost

Base: Number of SNPs per individual: 3,000,000

Costs: How much for a genome-wide SNP scan?

- Cost of 1 SNP: 0.30-0.45$ (soon 0.10-0.20$)
- Cost of a 10kb SNP map/individual: 90,000 (30,000)
- Cost of a 1000 individuals study: 90,000k (30,000k)
- Cost of 1000 complete maps: 900,000k (300,000k)

Time: How long does it take?

- 1 high throughput sequencer: 50,000 SNPs/day
- Effort 1000 10kb SNP maps: ~700 days/man
- Effort 1000 complete SNP maps: ~7000 days/man

Haplotypes

- LD (r2) distances can be used to identify haplotypes.
- Haplotypes are groups of SNPs transmitted in “blocks”.
- These blocks can be characterized by a subset of their SNPs (tags).
- Since they are the result of an underlying evolutionary process, they can be used to reconstruct ancestral DNA.

Figure removed due to copyright reasons.

The Importance of Haplotypes

- Haplotypes make a SNP map of the human genome redundant: as some SNPs will be transmitted together, we only need a subset of SNPs to tag the entire region.
- NHGRI launched in October the HapMap project: "*a description of the set of haplotype blocks and the SNPs that tag them. The HapMap will be valuable because it will reduce the number of SNPs required to examine the entire genome for association with a phenotype from all 10 million common SNPs to perhaps 200,000 to 300,000 htSNPs*."

Identifying Haplotypes

- Dely et at. report an high-resolution analysis of the haplotype structure of a stretch of chromosome 5q31 500Kbs long.
- There are 103 SNPs in the stretch.
- The SNPs were selected if the minor allele frequency was higher than 5%.
- Sample were 129 trios (nuclear families) of European descent with children affected by Crohn disease.
- Therefore, they had 258 transmitted and 258 non-transmitted chromosomes.

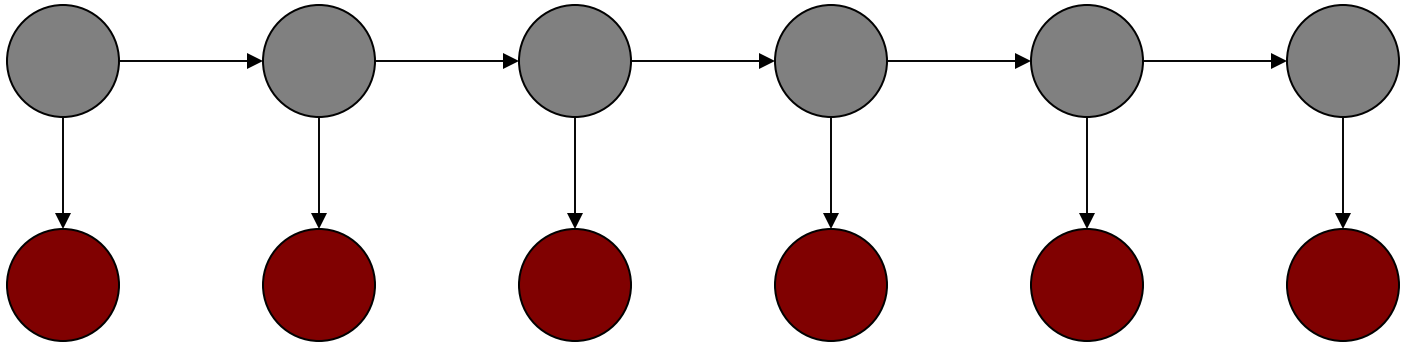
Haplotype Blocks

- The resulting picture portraits the stretch separated in 11 blocks separated by recombination points.
- Haplotype patterns travel together (blocks in LD) and therefore the authors infer 4 ancestral haplotypes.

Figure removed due to copyright reasons.

Hidden Markov Models

- To identify the 4 ancestral haplotypes, the authors use Hidden Markov Models (HHMs).
- The idea behind HMMs is that there is a directed process generating the data and that each observation is independent of the previous one given its generator and the previous generator.
- Each variable has 4 states (one per haplotype).



htSNPs Identification

- Johnson et al. propose a transmission-based method to identify Haplotype Tag SNPs (htSNPs), the necessary SNPs to identify an haplotype.
- The method is claimed to capture the *majority* (80%) of the haplotype diversity observed within a region.
- They genotyped polymorphisms in INS H19 SDF1 TCF8 and GAS2 in 418 UK families with at least 2 siblings with diagnosed type 1 diabetes.
- They constructed haplotypes at CASP8 CASP10 and CFLAR of 598 Finnish families with type 1 diabetes.

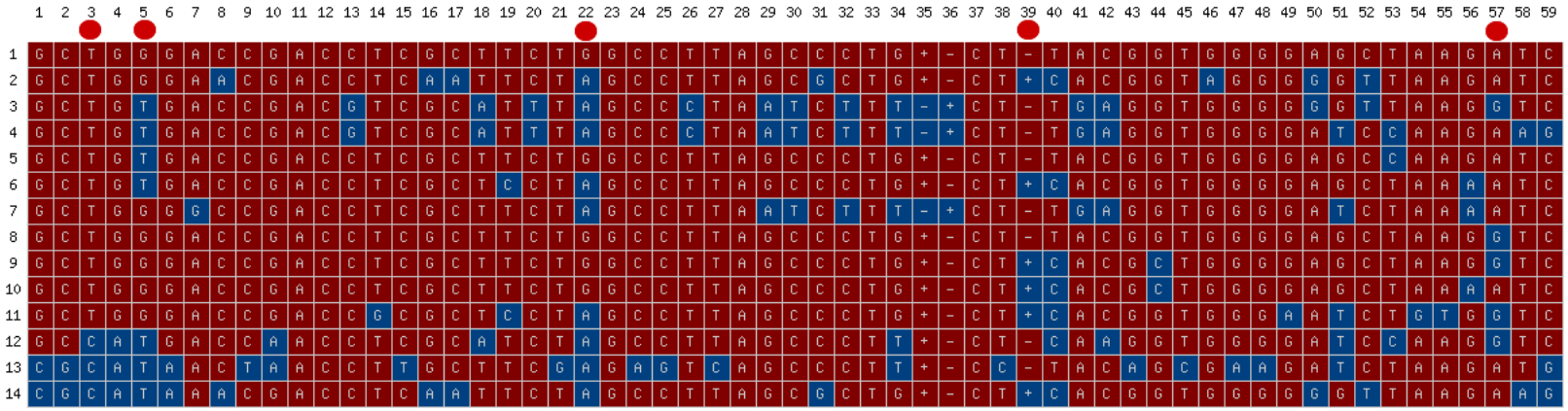
Haplotype Tagging

Haplotypes: As not all combinations appear, we need fewer SNPs.

Goal: Smallest set of SNPs deriving all the other SNPs.

htSNPs: These tagging SNPs are called haplotype tagging SNPs.

Problem: Intractable task (for 136 bases any relativistic machine would take more than the age of the universe).



TLR7

Intractable Solution

- The solution could be found by exploring, for each SNPs, all possible combinations of other SNPs.
- This is intractable: search space grows exponentially.
- Suppose to have a machine able to process a state in the time it takes light to cross a proton:
  - ✓ Light speed = 299792458 m/sec.
  - ✓ Size of a proton at rest = 0.7 fermi = 7E-16 m.
  - ✓ Time of light to cross a proton = 2.3E-24 sec.
  - ✓ Age of the Universe = 8 - 12,000,000,000 years.
  - ✓ Base per Age of the Universe = 136 (103 / year).

BEST (Determinist Tagging)

Goal: Smallest set of SNPs deriving all the other SNPs.

Condition: We assume the haplotype to be given.

Sufficient: A set of SNPs is sufficient if all the SNPs in the haplotype set can be derived from it.

Necessary: A set of SNPs is necessary if, removing any element, at least one SNP (including itself) is no longer derivable.

Minimal: The smallest set of necessary and sufficient SNPs from which all the SNPs are derivable.

htSNPs: members this set. Not necessarily unique.

With P Sebastiani

Central Principle

The fundamental property the method relies upon is:

Property: Every SNP derivable from a set of SNPs will be also derivable by any superset of such set.

Advantage: We do not need to know the set deriving a SNP to say that a SNP is derivable from a set.

T	G	A	G	T	T	A
G	T	T	C	T	A	A
G	T	A	G	A	T	A
T	T	A	G	A	A	T

Algorithm

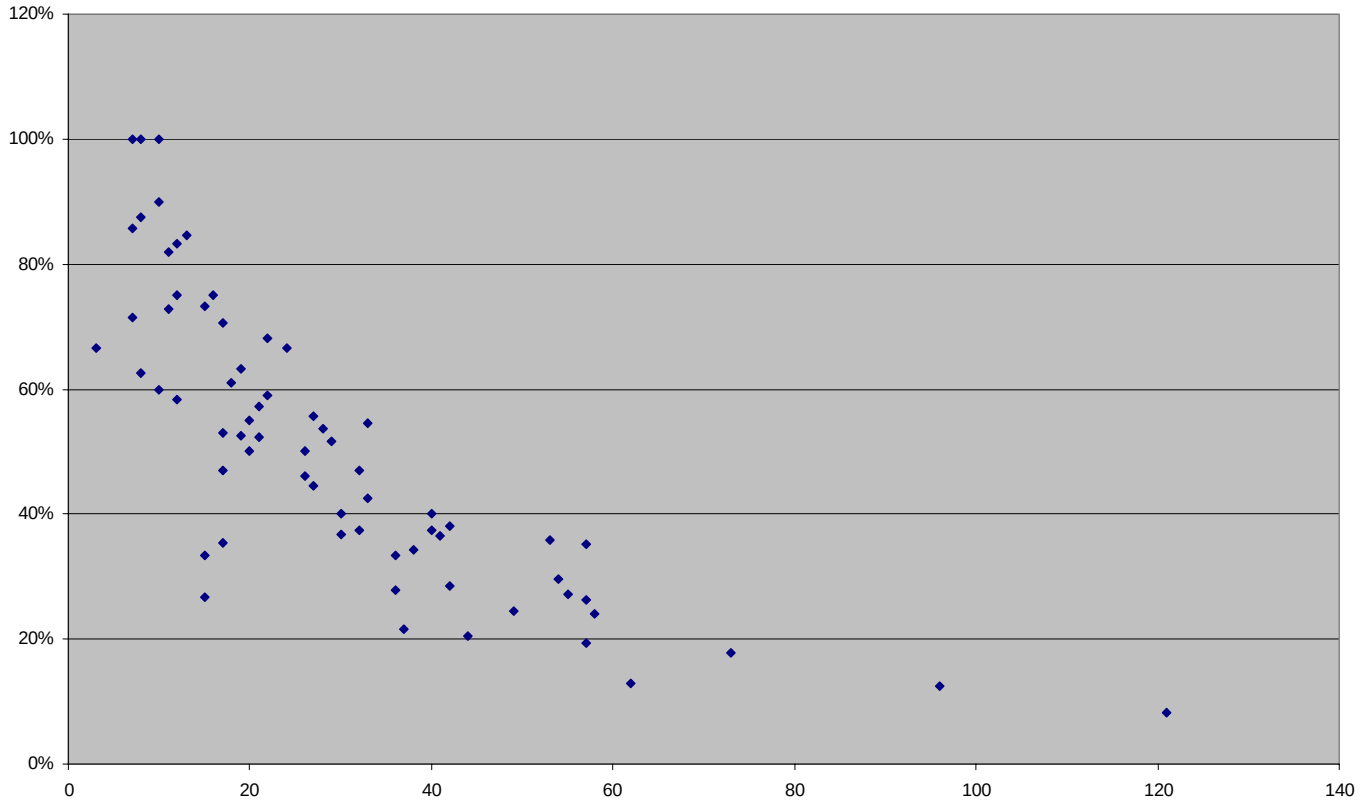
1. Turn the block into a binary table.
2. Identify those SNPs not derivable by all the others as members of the base (quadratic time).
3. Add to the base the SNP that allows the base to derive the maximum number of SNPs (cubic time).
4. If more than one SNP yields the maximum number of dSNPs, keep the alternative bases in parallel.
5. When the shortest base reaches sufficiency, stop.
6. If the base is sufficient and necessary, return it, else recursively call the procedure on the resulting base.

Advantages

- Provide a clear definition of htSNP.
- The htSNPs are necessary, sufficient and minimal.
- It is a deterministic process (optimal solution).
- We can tag any arbitrary region of the genome (such as a gene) rather than only haplotype blocks.
- The cost is a function of the resolution of the study.
- The notion of necessity sheds some light on the internal structure of the haplotype: a dSNP depends upon an htSNP if, when the htSNP is removed from the base, the dSNP is no longer derivable.

Savings

- htSNPs decrease as the size increases.



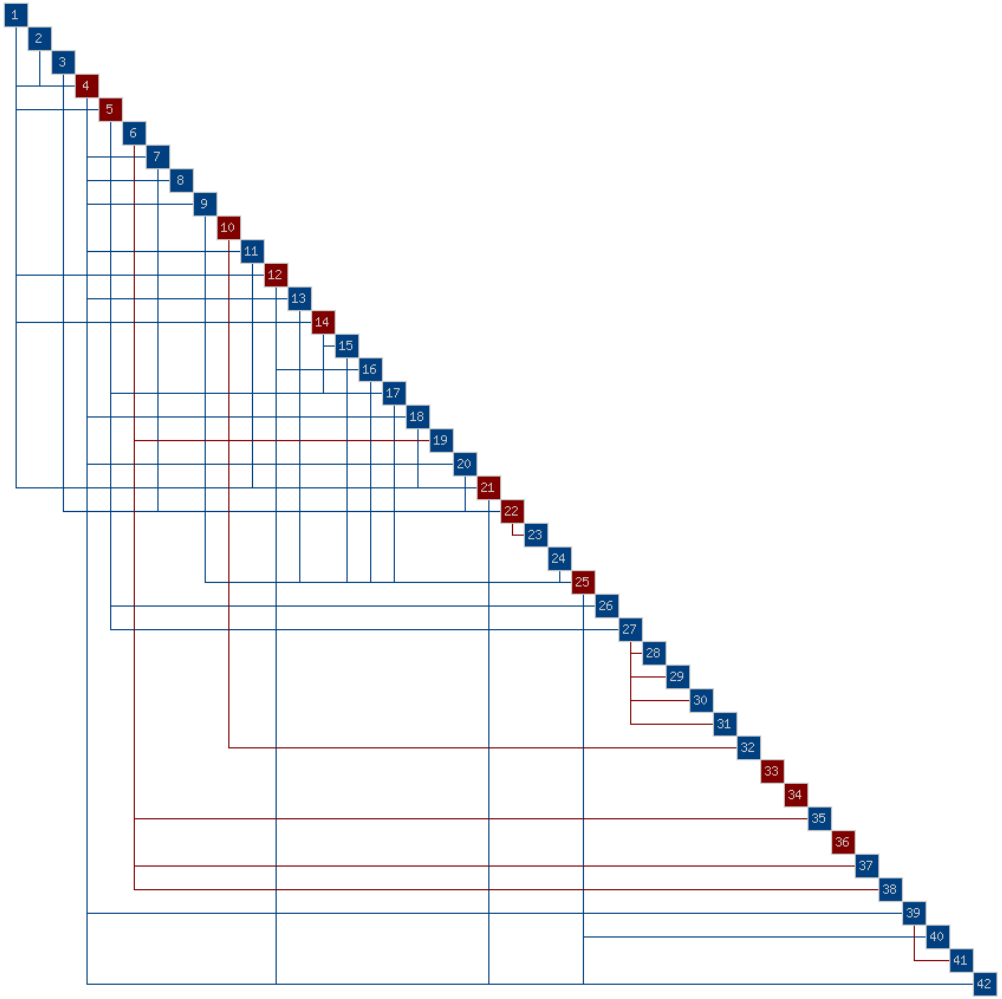
Tagging Haplotype Blocks

- Tagged the 11 haplotypes blocks found by Daly (2001).
- The results on the common haplotypes (with frequency threshold) are consistent with the ones we have found in the gene tagging exercise.
- If we tag the entire stretch for all haplotypes, the saving is much higher (46% vs 79%).
- Haplotype tagging makes block useless: good news because we need pedigrees.

Block	SNPs	Common Haplotypes		
		Haplotypes	*htSNPs*	*Ratio*
1	8	2	1	13%
2	5	2	1	20%
3	9	6	4	44%
4	11	4	3	27%
5	5	4	3	60%
6	5	4	3	60%
7	31	5	4	13%
8	7	6	4	57%
9	6	5	3	50%
10	7	6	4	57%
11	5	3	2	40%
Totals	**99**	**47**	**32**	**32%**

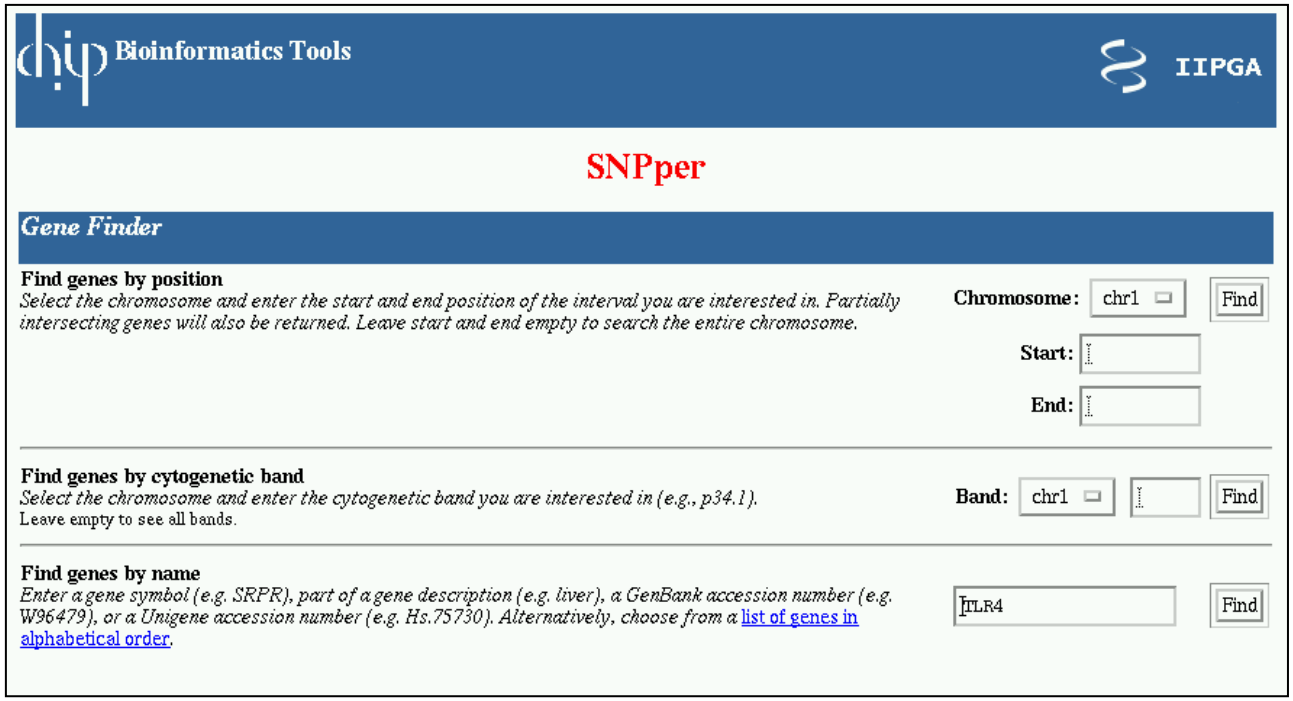
Block	SNPs	All Haplotypes		
		Haplotypes	*htSNPs*	*Ratio*
1	8	13	6	75%
2	5	7	4	80%
3	9	21	8	89%
4	11	19	8	73%
5	5	11	5	100%
6	5	11	5	100%
7	31	52	19	61%
8	7	17	7	100%
9	6	15	5	83%
10	7	21	7	100%
11	5	8	4	80%
Totals	**99**	**195**	**78**	**79%**
As a Single Block				
(1 -11)	103	359	47	46%

Dependencies



Annotating SNPs

- SNPper is a program for retrieval of SNPs information from public databases (eg dbSNP) developed by A. Riva (http://bio.chip.org/biotools).



Courtesy of Dr. Alberto A. Riva. Used with permission.

Characterizing Phenotypes

Simple phenotypes: Mendelian diseases usually have also the advantage of simple (binary) phenotypes.

Complex diseases: Twenty years of AI in medicine show that often diseases do not obey these patterns.

Dissecting phenotypes: It is critical as dissecting gene expression patterns.

Animal models: QTL strategy may be an answer.

Opportunity: Better clinical definition of disease states.

Clinical data: With dropping costs of sequencing, good clinical data about patients are the real wealth.

Take Home Messages

The Past:

Hypothesis-driven phenotype-focused data;
Interesting discoveries with simple models.

The Present:

HGP changes the perspectives of genetic studies;
SNPs are a critical tool to break the code;
HGP technology makes sequencing a commodity.

The Future:

Exploratory genotyping will streamline discoveries;
Phenotypes will be the real goodies of the future;
The challenge is to handle complex traits.