SNPs and Haplotypes

Bioinformatics
NSF

SNPs Concepts

• SNPs – what are they?
• Why are SNPs important?
• SNPs and linkage disequilibrium
• Common SNPs / haplotype blocks
• SNPs / Haplotype block – navigation
• Building complex traits and ‘the myth of race?’, the role of SNPs and haplotypes

SNPs - Introduction

• Single Nucleotid Polymorphism
• Occurs once in human evolution
• Estimate of 1 bp in 600 – 1000 bp
• Occur mostly in introns (2/3)
  – Regulatory regions, leading to cancer
• Often lethal when in exons (1/3)
  – Leading to a fatal amino acid substitution

What are Single Nucleotide Polymorphisms (SNPs)?

ATGGTAAGCTGAGCTGACTTAGCTTACGT-AT
ATGGTTAAACCTGATGCTGACTTACGTCAT
↑ ↑ ↑
SNP SNP indel

SNPs result from replication errors and DNA damage
They are a ‘polymorphic’ bit state at a nucleoside address

What (exactly) is a SNP?

• A SNP is defined as a single base change in a DNA sequence that occurs in a significant proportion (more than 1 percent) of a large population.
• Occurs exactly once in human evolution
• Degenerate ‘bit state’ at a genomic address (A, T, C, or G) usually n=2
• Degenerate ‘bit state’ = polymorphism

SNPs / Polymorphisms

• A Single Nucleotide Polymorphism is a source of variance in a genome. A SNP (“snip”) is a single base mutation in DNA.
• SNPs are ‘conserved’ across the genome, often in patterns called ‘haplotype blocks’
• SNPs are the most simple form and most common source of genetic polymorphism in the human genome (90% of all human DNA polymorphisms are associated with SNPs).
Why are ‘SNP’ Polymorphisms Useful?

- It’s sometimes possible to correlate a SNP with a particular trait or disease.
  - This is known as association genetics.
- Susceptibility to disease may also be described as an ‘unfortunate trait’.
- Traits are also ‘larger’ than genes.
- SNPs in (regulatory) intragenic regions may be as important as (coding) exons.

SNP Applications in Medicine

- Gene discovery and allele mapping
- Association-based (drug) candidate
  - polymorphism testing of a trait pool
- Diagnostics / risk profiling
- Drug response prediction
- Homogeneity testing / study design
- Gene function identification

Genetic Polymorphism

- Genetic Polymorphism: A difference in DNA sequence among individuals, groups, or populations. **One or more SNPs.**
- Genetic Mutation: A change in the nucleotide sequence of a DNA molecule. Genetic mutations are one type of genetic polymorphism (but less than 1%).
- Polymorphism is common, mutation rare
  - Polymorphism is the ‘stuff of variation’

Coding Region SNPs

- Types of coding region SNPs:
  - 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.
  - One half of all coding sequence SNPs result in non-synonymous codon changes. (but half do)
SNPs and Protein Structure

Types of SNPs

- There are two types of nucleotide base substitutions resulting in SNPs:
  - Transition: substitution between purines (A, G) or between pyrimidines (C, T).
    Constitute two thirds of all SNPs.
  - Transversion: substitution between a purine and a pyrimidine.

How Common are SNPs?

- Of the roughly 5 to 6 million SNPs
  - Half appear in 50% of the population
  - One quarter in 25% of the population
  - One eighth in 12.5% of the population
- The pattern of inheritance of SNPs appears to follow the pattern of inheritance of haplotypes (linkage)

Sequence Variation in Humans

- Population size: 6x10^8 (diploid)
- Mutation rate: 2x10^-3 per bp per generation
- Expected "hits": 240 (humans with a SNP per bp)
  - Every variant compatible with life exists in the population BUT most are vanishingly rare
- Compare 2 haploid genomes: 1 SNP per 1331 bp*


Strategies to Find SNPs

- Mine them from existing genome resources
- Targeted SNP discovery in candidate genes
- HapMap study to define global SNP population
- CardioGenomics - http://www.cardiogenomics.org/
- InnateImmunity - http://innateimmunity.net/
- SeattleSNPs - http://pga.mbl.washington.edu/
- Southwestern - http://pga.swmed.edu/
- Perlegen website – http://www.perlegen.com/
- SNP Consortium website - http://snp.cshl.org/
Sequence-Based Detection and Genotyping of SNPs

- Jim Sloan, Tushar Bhangle (PolyPhred)
- Matthew Stephens, Paul Schuel (Quality Scores for SNPs)
- PS Green, Brent Ewing, David Gordon (Freez, Phrep, Consed)

SNP Discovery and Genotyping Workshop

SNPs and Variation

- In human beings, 99.9 percent of bases are same
- Remaining 0.1 percent makes a person unique.
  - Different attributes / characteristics / traits
  - how a person looks,
  - diseases he or she develops.
- These variations can be:
  - Harmless (change in phenotype)
  - Harmful (diabetes, cancer, heart disease, Huntington's disease, and hemophilia)
  - Latent (variations found in coding and regulatory regions, are not harmful on their own, and the change in each gene only becomes apparent under certain conditions e.g. susceptibility to lung cancer)

Why Create SNP Profiles?

- Genome of each individual contains a distinct SNP pattern (haplotype block).
- People can be grouped based on their SNP profiles (association studies).
- SNP profiles may be important for identifying response to drug therapy.
- Correlations might emerge between certain SNP profiles and specific responses to treatment (good and bad).

Populations Based on SNPs

- Single Nucleotide Polymorphisms (SNPs), Haplotypes, Linkage Disequilibria, and the Human Genome

Techniques to Detect Known Polymorphisms

- Hybridization Techniques
  - Microarrays
  - Real time PCR
  - HTS SNP arrays
- Enzyme based Techniques
  - Nucleotide extension
  - Cleavage
  - Ligation
  - Reaction product detection and display
- Comparison of SNP Assay Techniques

SNP Genotyping

- SNPs can be measured using High Throughput Screening with custom (HTS) microarray technology (Applied Biosystems)

SNP Discovery and Genotyping Workshop
Techniques to Detect Unknown Polymorphisms

- Direct Sequencing
  - HTS SNP sequencing
- SNP Microarray
  - Rapid SNP genotyping
- Cleavage / Ligation
- Electrophoretic mobility assays
- Comparison of Techniques used to detect unknown (SNP) polymorphisms

More SNP Terminology

- Polymorphism / Haplotype
  - Correlation of characters states among polymorphic sites (across the genome). SNP patterns = "blocks"
  - Insufficient passage of time to randomize character states by meiotic recombination – patterns conserved
- Haplotypes are ‘blocks’ of associated SNPs
- Haplotypes may be ‘too recent’ ...
  - Not enough time for recombination to merge SNPs
  - Or SNP recombination may be a more difficult process

SNPs and Haplotypes

- SNP: Single Nucleotide Polymorphism
- Haplotype: A set of closely linked genetic markers present on one chromosome which tend to be inherited together (not easily separable by recombination).

Set of SNP polymorphisms: a SNP haplotype

From SNP to Haplotype

Phenotype | DNA Sequence | Haplotypes
--- | --- | ---
Black eye | GATATTGATACGGA-T | AG 2/6
Brown eye | GATATTGATACGGA-T | GTA 3/6
Black eye | GATATTGATACGGA-T | AGA 1/6
Blue eye | GATATTGATACGGA-T |
Brown eye | GATATTGATACGGA-T |
Brown eye | GATATTGATACGGA-T |

Seoul National University School of Public Health

Seoul National University School of Public Health
SNPs and Linkage Disequilibrium (LD)

- Recent Resurgence of LD Study Motivated by SNP Markers / haplotypes (HapMap project)
- High density: ~ one SNP in every 600 bp in the human genome make them easy to map.
- Simple SNPs: Biallelic (occur on both alleles)
- Common: ~93% are found globally (among human populations); ~7% are restricted to local populations. (NHGRI, 2001)

SNPs / Linkage Disequilibrium

- Generally speaking…
- SNPs should be inherited independently
  – Following Mendelian inheritance
- Many SNPs appear to be co-inherited
  – Creating ‘hot spots’ in the human genome
- Blocks of SNPs – ‘SNP haplotypes’
  – We don’t why or how, but they exist

Linkage Disequilibrium

Haplotype is the pattern of alleles on a single chromosome
- 4 possible haplotypes
Linkage Disequilibrium (LD) describes the allelic association between two SNPs
Two popular LD statistics:
\[ D' \]
\[ r^2 \]

Complete LD

Unequal allele frequency
Allelic association is as strong as possible
- 2 haplotypes observed
- No detected recombination between SNPs
- Genotype is not perfectly correlated
\[ D' = 1 \]
\[ r^2 < 1 \]

Perfect LD

Equal allele frequency
Allelic association is as strong as possible
- 2 haplotypes observed
- No detected recombination between SNPs
- Genotype is perfectly correlated
\[ D' = 1 \]
\[ r^2 = 1 \]

Rational SNP Selection

- Select SNPs to genotype on the basis of LD
- Some SNPs are in LD with many other SNPs
- Some SNPs are in LD with no other SNPs
- SNPs between a pair of associated SNPs are not necessarily associated with the flanking SNPs
Haplotypes – Understanding the ‘Nature’ of SNPs

- Visualizing haplotype blocks
- Haplotypes and SNP distribution
- Haplotypes as genomic ‘barcodes’
- HapMap Project – a global SNP study
- Perlegen Haploype / Genomic browsers
- Haplotypes and traits, mapping *alleles*
- The myth of *race* – insights on *variation*

The International HapMap Project, Nature 2003

Haplotypes – SNP Barcodes

Visualizing Haplotype Blocks

http://acg.media.mit.edu/people/fry/haplotypes/
SNP HapMap Project
http://www.HapMap.org/

• Sequence genomes of a large number of people (n >10,000 ethnically diverse)
• Compare the base sequences to discover SNPs, their location, and their frequency.
• Generate a single map of the human genome containing all possible SNPs => SNP maps

Perlegen Haplotype Study

*Global patterns of human DNA sequence variation (haplotypes) defined by common single nucleotide polymorphisms (SNPs) have important implications for identifying disease associations and human traits. We have used high-density oligonucleotide arrays, in combination with somatic cell genetics, to identify a large fraction of all common human chromosome 21 SNPs and to directly observe the haplotype structure defined by these SNPs. This structure reveals blocks of limited haplotype diversity in which more than 80% of a global human sample can typically be characterized by only three common haplotypes.*

Abstract from seminal article by Cox Science 294: 1719-1723 (2001)

Haplotypes in the Genome

• Defined as a pattern of SNPs that appears as an ‘associated block’ on one or more chromosomes
• There are (estimated*) to be roughly 200K to 300K haplotypes in genome
• Of these, most can be defined by the identity of 3 SNPs, and always < 10

* Cox et.al. Perlegen Haplotyping of chromosome 21

Golden Path / Human Variation

• 3 billion base pairs
• 6 to 10 million SNPs
• 200K – 300K interrelated haplotype ‘blocks’ in the human population
• Each block is about 7.8 K bp – (this is an average)
• Each block contains roughly 10 SNPs – (of which 1 to 3 define the haplotype)

Haplotype Questions

• David Cox – traits vs. genes?
• Disequilibrium – patterns of inheritance?
• ‘Hot spots’ – defined by haplotypes?
• Are SNPs more definable than ‘race’?
  – One race, 200,000 ‘haplotype / traits’
  – One race, common haplotype patterns?
• How and why do haplotypes occur?
  – Is there a benefit to SNP patterns?
SNPs per Haplotype Block
- In common haplotypes (>80%) 3 SNPs can determine the identity of the block, and often just 1 SNP determines the haplotype identity (3 SNPs / haplotype)
- At most, only 10 SNPs (10% of roughly 100 SNPs) will define the 'identity' for the majority of observed haplotypes
- These SNPs are often called 'markers'

Identification
- In half of all haplotype blocks, 3 or fewer SNPs are needed for an identification
- In haplotype blocks with 3 SNPs, only one SNP is needed to identify a block
- In other cases, 10 SNPs define a block
- In all cases, 10% SNP identification is enough to create a haplotype definition

Haplotyping Tools (Now/Later)
- High density nucleotide arrays
  - Rapid SNP genotyping
- High resolution maps of chromosome
  - Data mining tools for pattern recognition
- Key word / ontologies for disease, traits, other definitions of human variability
- Eventually haplotype expression data

Haplotype Utility
- Are SNP patterns as important as SNPs in characterizing disease susceptibility?
- Are haplotype patterns a better way to define an individual's (UID) genome?
- Is this a better tool to understand the 'evolution of race', or 'myth of race'?
- Is this a method to identify proteomes?
  - Proteome = (UID) genome * expression?

Haplotype Exercise Overview
- Perlegen haplotype browser
- Database of haplotypes
- Navigate along a chromosome
- Pull up a haplotype block
- Examine blocks and SNPs
- Follow the SNP links into NCBI
- Perlegen genotype browser

Haplotype Exercise (I)
- Go to Perlegen haplotype browser
- Download the paper by David R Cox
  
- Scan the paper, which is a tough to read (technically), and use the next slides as a review of the high points
- The image of haplotype blocks is key
  - (This was shown earlier in the presentation)
Haplotype Exercise (II)

- Click in the middle of 21q21+2 (or anywhere along the chromosome)
- That will bring up a haplotype block
- Scroll down that page, and look at the entries for each of the rows in the block
- At the bottom of the page, look at the base calls for each (block row) entry
- Follow the bottom links to NCBI dbSNP
Perlegen Cox Paper (I)

• 24 separate ethnically diverse individuals
• 35,989 SNPs were identified on chromosome 21 (32,397,439 bp)
• 47% of 53,000 common SNPS occur with an allele frequency of 10%
  – 24,000 SNPs are highly reproducible from human to human - polymorphism

Perlegen Cox Paper (II)

• Used the set of 24,000 SNPs (with a minor allele) to define a haplotype
• Blocks define haplotype patterns
• The four most common haplotypes account for 80% of all occurrences
  – These ‘universal haplotypes’ defined by an address in the genome with SNP variation
  – Their distribution follows a ‘normal curve’

Perlegen Cox Paper (III)

• Alleles making up blocks of such SNPs are often correlated – resulting in reduced genetic variability and defining a limited number of ‘SNP haplotypes’.
• Leads to ‘linkage disequilibrium’
• 80% of the haplotype structure is defined by < 10% of the SNPs in a haplotype block. This is universal.

Perlegen Cox Paper (IV)

• Most common haplotype pattern (of the four universal haplotypes) is found in:
  • 1st – 50% of all individuals
  • 2nd – 25% of all individuals
  • 3rd – 12.5 % of all individuals
  • 4th – ~ 6% of all individuals (estimate)
• These total 93% of all human (genomes)

Perlegen Cox Paper (V)

• Average size of a block is 7.8 Kbytes
• Some blocks are up to 114 SNPs and 115 Kbytes (1 SNP per 1000 bp)
• 10% of 114 SNPs or only 11 SNPs are needed to define the largest haplotypes
• Average distance among all SNPs was 900 bp, and among 24,000 common SNPs, was 1,300 bp
Perlegen Cox Paper (VI)

- On average, there are 2.7 common haplotypes per block, defined as haplotypes observed on multiple chromosomes
- 94% of blocks contained fewer than 3 SNPs
- Exonic bases:
  - > 10 SNPs
  - 3 to 10 SNPs
  - < 3 SNPs

Perlegen Cox Paper (VII)

- Haplotype blocks are defined by their genetic information, and not on knowledge of how this information originated, or why it exists.
- Perlegen (David Cox) suggests that perhaps ‘human traits’ associated with haplotypes are a better way to approach genetic variation in the human genome

The Myth of Race?

- 93% of SNPs are ‘universal’ - ‘public SNPs’
- 80% of SNP variation defined by four or less haplotypes within a haplotype block
- Haplotype patterns and ‘multigenic’, meaning that blocks can span genes (CDS and introns)
- Traits may be multigenic and multi-block.
- We are more complex as individuals than we are differentiated by race, but ‘race’ can be studied

Perlegen Genome Browser

- Database of haplotypes / SNPs and LD
- Data separated by ‘race’
  - European
  - African American
  - Han Chinese
- Navigate along a chromosome
- Examine SNP frequencies by ‘race’
  - Study alleles and population distributions
  - http://genome.perlegen.com/browser/
Whole-Genome Patterns of Common DNA Variation in Three Human Populations

David A. Hinds¹, Laura L. Stone¹, Geoffrey S. Nisela,²
Brian Halperin³, Susan E. Eichten,⁴ Dennis G. Ballinger,⁵
Kelly A. Freimer⁶, David R. Cox⁵

Individual differences in DNA sequence are the genetic basis of human variability. We have characterized whole-genome patterns of common human DNA variation by genotyping 1,500,000 single-nucleotide polymorphisms (SNPs) in 75 Americans of European, African, and Asian ancestry. Our results indicate that these SNPs capture most common genetic variation as a result of large-scale admixture and the isolation among distinct SNP alleles. We observe a strong correlation between extended regions of linkage disequilibrium that are larger than 100 kilobases in length, as well as strong evidence of an excess of SNPs in the noncoding regions of the human genome. Linkage-disequilibrium analyses for these polymorphisms can provide insights into the evolutionary history of the human species and the mechanisms of selective sweeps in the human population. We are continuing to identify SNPs that are associated with complex human traits and diseases by using the patterns of linkage disequilibrium to search for genetic variations that are linked to specific conditions.
Perlegen Next Steps

- Identify SNP patterns in thousands
- Map SNPs / haplotypes to disease
- Map SNPs / haplotypes to human ‘traits’
- Association maps for haplotypes
  - How do blocks interrelate?
- Create a ‘Hap Map’ for human genome
- Ability to survey the entire genome will dramatically increase the power of genetic association analysis.

Conclusions

- SNPs (single nucleotide polymorphisms) are abundant and useful genetic markers
  - Disease, drug resistance or adverse effects, etc
- Linkage disequilibrium describes the tendency of many SNPs to be inherited in patterns or blocks
  - Typical blocks include 7 to 8 common SNPs / 7.8 Kb
- These haplotype patterns, or blocks, appear in >90% of the population, in over 250,000 identified blocks
  - Often 3 or fewer SNPs can define a haplotype block
- Association studies of individuals and populations that show disease or drug response behaviour in progress
  - Where we’ll see the benefit of haplotype research!

Further Reading / Resources

- Applied Biosystems: http://www.appliedbiosystems.com/
- Perlegen: http://www.perlegen.com/
- SNP Consortium: http://snp.cshl.org/
- International SNP Working Group Data (Nature)
- Google ‘haplotypes, SNPs, haplotype maps, SNP genotyping, linkage disequilibrium’, HapMap, and all (word / keyword) combinations of the above

Presentation References

- Introduction to SNPs: Discovery of Markers of Disease
- SNP Mapping using Genotypes with Unique Sequences
- The Structure of Haplotype Blocks in Human Genome
- Using haplotype blocks to map human complex traits
- High Resolution haplotype structure in human genome
- Detection of regulatory variation in mouse genes
- Resolution of Haplotype and Haplotype Frequencies from SNP Genotype of Pooled Samples
- Visualizing Haplotype Blocks
- Visualizing Haplotype Blocks
- Visualizing Haplotype Blocks
- Visualizing Haplotype Blocks
- Visualizing Haplotype Blocks
- Visualizing Haplotype Blocks
- Visualizing Haplotype Blocks

Presentation Acknowledgements

1. Single Nucleotide Polymorphisms (SNPs), Haplotypes, Linkage Disequilibrium, and the Human Genome Manish Anand, Nihar Sheth Jim Costello
2. High-resolution haplotype structure in the human genome Mark J. Daly, John D. Rioux, Stephen F. Schaffner, Thomas J. Hudson & Eric S. Lander
4. Selecting SNPs for Genotype-Phenotype Analysis Using Allelic Association Linkage Disequilibrium) Christopher Carlson csc47@u.washington.edu

Copyright Status

- Slides used in this presentation include material from several presentations downloaded from Programs for Genomic Applications Educational Activities. This material is used consistent with the educational license.
- Web citations / links are used whenever possible (see the presentation acknowledgements slide).
- Screen grabs from the Perlegen website are included in the tutorial section on the Perlegen haplotype browser.
- Inquiries for educational use of this material should be forwarded to me at rdcormia@earthlink.net Thank yeB!