Two Standard Categories of NGS Assemblers

- “Overlap-layout-consensus” (OLC) approach
  - Overlap graph

- de Bruijn graph (DBG) approach
  - k-mer graph
  - Especially useful for assembly from short reads

A simple case

- Let’s consider the following sequence:
  
  Generate k-mers 
  \( k = 3 \)

  Genome: ATGCGGTGCAATG

  A circular genome

  For each k-mer, let’s create the (k-1)-mer “prefix” and “suffix” nodes
CSE 427 Computational Biology, Winter 2015

A simple case

- Let's consider the following sequence:

Genome: ATGCGGTGC

---

A less simple case

(a) [5 points] Draw the de Bruijn graph for the following set of reads:

ATTAC TACAG GATTA ACAGA CATAC ATCA AGATT

---

Why are we so different?

- Different "phenotypes" (observable traits)
  - Appearance
  - Disease susceptibility
  - Drug responses

- Different "genotype"
  - Individual-specific DNA
  - 3 billion-long string

---

Genetics
Motivation

- Which sequence variation affects a phenotype?
  - Better understanding disease mechanisms
  - Personalized medicine

Sequence variations

DNA is 3 billion long!

Instruction

Obese? 15%
Bald? 30%
Diabetes? 6.2%
Parkinson’s disease? 0.3%
Heart disease? 20.1%
Colon cancer? 6.5%

Outline

- Basic concepts
  - Meiosis, genetic recombination
  - Allele, allele frequencies, genotype frequencies
  - Genotyping

- Statistical methods for mapping QTL
  - What is QTL?
  - Experimental animals
  - Analysis of variance
  - Statistical significance of the LOD score

Humans have 23 pairs of chromosomes

Source: http://www.dirkschweitzer.net/
Homologous chromosomes

Transmission of genetic information

- **Meiosis** is the process that results in the formation of gametes (sperm cells and egg cells)

Forming the next generation

Outline

- Basic concepts
  - Meiosis, genetic recombination
  - Allele, allele frequencies, genotype frequencies
  - Genotyping

- Statistical methods for mapping QTL
  - What is QTL?
  - Experimental animals
  - Analysis of variance
  - Statistical significance of the LOD score
Alleles

- Alternative forms of a particular sequence
- Each allele has a frequency, which is the proportion of chromosomes of that type in the population

Allele frequencies for C, G, –

Genotype

- The pair of alleles carried by an individual
  - If there are $n$ alternative alleles ...
  - ... there will be $n(n+1)/2$ possible genotypes
  - In most cases, there are 3 possible genotypes

Homozygotes
- The two alleles are in the same state
  - (e.g. CC, GG, AA)

Heterozygotes
- The two alleles are different
  - (e.g. CG, AC)

Genotype frequency

- Since alleles occur in pairs, these are a useful descriptor of genetic data.

- However, in any non-trivial study we might have a lot of frequencies to estimate.
  - $p_{AA}$, $p_{AB}$, $p_{AC}$, $p_{BB}$, $p_{BC}$, $p_{CC}$, ...
Genotype vs. allele frequencies

- Genotype frequencies lead to allele frequencies.
- For example, for two alleles:
  - \( p_A = p_{AA} + \frac{1}{2} p_{AB} \)
  - \( p_B = p_{BB} + \frac{1}{2} p_{AB} \)
- However, the reverse is also possible!

Hardy-Weinberg Equilibrium

- Relationship described in 1908
  - Hardy, British mathematician
  - Weinberg, German physician
- Shows \( n \) allele frequencies determine \( \frac{n(n+1)}{2} \) genotype frequencies
  - Large populations
- Random union of the two gametes produced by two individuals

HWE assumption

- Allele frequencies lead to genotype frequencies.
- For example, for two alleles:
  - \( p_{AA} = p_A^2 \)
  - \( p_{BB} = p_B^2 \)
  - \( p_{AB} = 2 p_A p_B \)

Outline

- Basic concepts
  - Meiosis, genetic recombination
  - Allele, allele frequencies, genotype frequencies
  - Genotyping
- Statistical methods for mapping QTL
  - What is QTL?
  - Experimental animals
  - Analysis of variance
  - Statistical significance of the LOD score
Genotyping chip

Probes

![Genotyping chip image](image)

### Outline

- **Basic concepts**
  - Meiosis, genetic recombination
  - Allele, allele frequencies, genotype frequencies
  - Genotyping

- **Statistical methods for mapping QTL**
  - What is QTL?
  - Experimental animals
  - Analysis of variance
  - Statistical significance of the LOD score
Definition of QTLs
- The genomic regions that contribute to variation in a quantitative phenotype (e.g. blood pressure, height)

QTL analysis
- Linking phenotype data and genotype data (genetic markers), in order to explain the genetic basis of variation in complex phenotypes

Goals:
- Link certain complex phenotypes to specific regions of chromosomes
- Identify precise location of these regions and the interaction with phenotypes

Quantitative Trait Locus (QTL)
- Two or more strains of organisms that differ genetically with regard to the phenotype of interest
  - Experimental animals
    - Backcross experiment (only 2 genotypes for all genes)
    - F2 intercross experiment
  - Genetic markers that distinguish these parental lines
    - Single nucleotide polymorphisms (SNPs)

Backcross experiment
- Inbred strains
  - Homozygous genomes
- Advantage
  - Only two genotypes
- Disadvantage
  - Relatively less genetic diversity

F2 intercross experiment
- Parental generation
- First filial (F1) generation
- F2 generation

Karl Broman, Review of statistical methods for QTL mapping in experimental crosses
Trait distributions: a classical view

QTL analysis

- **Goals**
  - Identify the genomic regions (QTLs) contributing to variation in the phenotype.
  - Identify at least one QTL.
  - Form confidence interval for QTL location.
  - Estimate QTL effects.

Outline

- **Basic concepts**
  - Meiosis, genetic recombination
  - Allele, allele frequencies, genotype frequencies
  - Genotyping

- **Statistical methods for mapping QTL**
  - What is QTL?
  - Experimental animals
  - Analysis of variance
  - Statistical significance of the LOD score