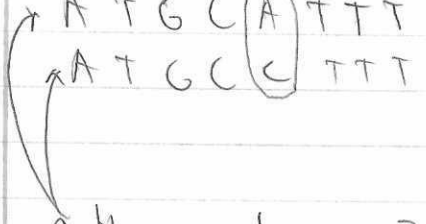
GDA II, §1

- This book is about analyzing DNA data



↑ Humans have 2 copies
↑ Some plants have more than 2
↑ Most loci are the same in the same species
↑ this individual has one difference

- "SNP" = "Single Nucleotide Polymorphism"
  = Locus on genome where there is a single difference that varies in the population

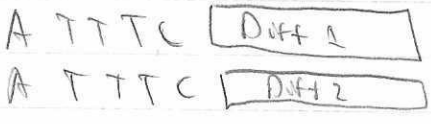
- "Allele" = a difference
  ↑ A allele and C allele above

- There are many ways DNA can differ at a locus

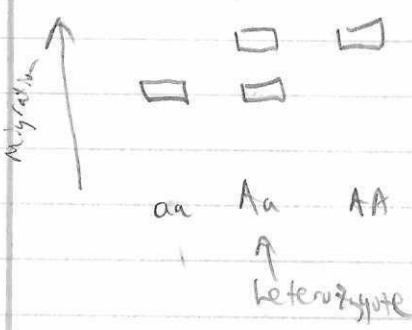
- INDEL (insertion/deletion)

  A T T G C T ← Insert here.
  A T T _ C T ← Delete here.

- Large structural changes



- This first chapter describes assays to see these DNA differences.
- Phenotype: observed consequence of an allele
  - Round peas versus wrinkled peas
- Homozygote: 2 of the same allele
- Heterozygote: Contains different alleles.
- "Dominant Phenotypes" – AA observe round
  Aa observe round
  aa observe wrinkled ← "recessive"
  ↑ only expressed if homozygote
  ↑ expressed if homozygote or heterozygote
- "Codominant" – can see presence/absence of allele no matter what
- "Allozyme" – old fashioned DNA → RNA → protein
  - how far a protein migrates
  - changes based on DNA sequence



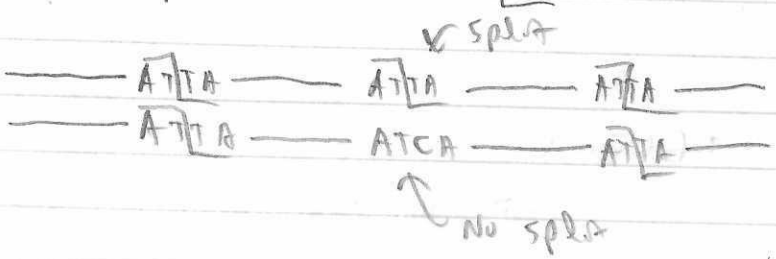
- Proteins: Sequences of 20 Amino Acids
  I can sequence these

Sometimes still used

- RFLP (Restriction fragment length Polymorphisms)

- Restriction enzymes: Proteins that split DNA wherever they find a certain sequence

- Ex. Split at $-\overline{ATTA}-$



- So set of fragment sizes indicates presence/absence of allel

- Steps: 1.) Apply enzyme to DNA
  2.) Let migrate in gel
  3.) smaller molecules migrate (further) differently
  4.) Use a probe to detect region of interest
  ↳ STR gives you presence/absence

Sometimes still used

- Microsatellites = short tandem Repeats (STR)
  = simple sequence repeats (SSR)
  = region of DNA that has lots of repeats

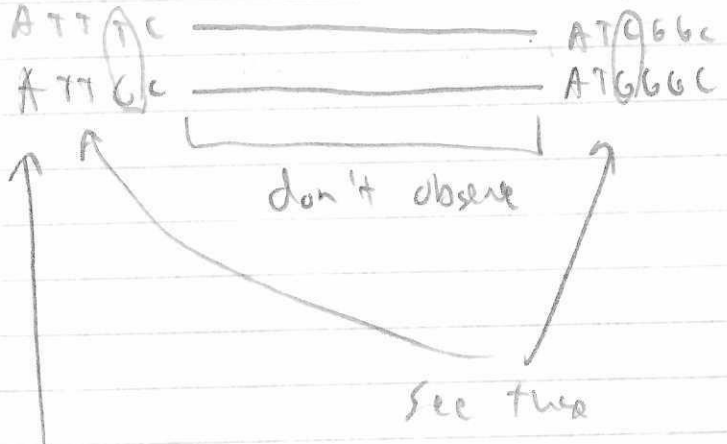
ATC ATC ATC ATC ...

↳ Length of this repeat is an allele

- All of these methods try to find properties of DNA,
- If we can see DNA, don't need them
- DNA sequencing is possible now and is gold standard
  - but is sometimes really noisy, so sometimes older technologies are used
- Sequencing data:

  ATTGCATT
  ATTGCTTT
  ↑ SNP, we see it!

- Sometimes, have sequences of different regions of genome

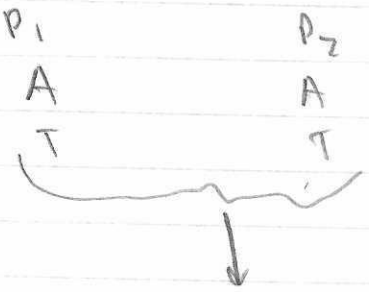


Is T or sub molecule as C or as G?
I can't tell sometimes

Genetic vs Statistical Sampling:

- Statistical Sampling: There is a population of size N, but we only observe n of them

- Genetic Sampling: From generation to generation, random alleles will segregate to offspring



F_1	AA	AT	TT
Prob:	1/4	1/2	1/4

↳ This shows up in later chapters (not 2 and 3)
↳ we are mostly interested in present population

- Notation:

N: Population size
n: Sample size
A: reference allele
a: alternative allele
If more than 2 alleles, $A_1, A_2, \ldots A_6$
p_A: "Allele frequency"
↳ Proportion of A's in the <u>population</u>

- If 2 loci, alleles are A/a and B/b

P_{AB} = proportion of genomes w/ allele A at first locus and allele B at second locus.

- $\underline{P}_{AA}$ = genotypic frequencies

↑
capital P

$\underline{P}_{AA}$ = proportion 2 A's @ a locus
$\underline{P}_{Aa}$ = proportion 1 A and 1 a @ a locus
$\underline{P}_{aa}$ = proportion 2 a's @ a locus

- $\underline{P}_{ab}^{AB}$ = proportion AB on one chromosome and ab on other

$\underline{P}_{aB}^{Ab}$ = proportion Ab on one chromosome and aB on other

$\underline{P}_{AaBb}$ = proportion w/ 1 A and 1 a at locus 1
and 1 B and 1 b at locus 2
I but don't know whether A pairs w/ B or b

N_{AA} = # AA genotypes
N_{Aa} = # Aa genotypes
N_{aa} = # aa genotypes

$N = N_{AA} + N_{Aa} + N_{aa}$ = total # people

$N_A = 2N_{AA} + N_{Aa}$ = # A genomes

- Sample frequencies w/ tilde's

$\tilde{p}_A = \frac{1}{2n} \wedge_A$

$\tilde{P}_{AA} = \frac{1}{n} \wedge_{AA}$

- Estimates of parameters have hats.

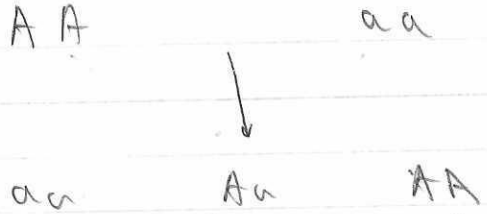
$\hat{p}_A = \hat{p}_A$

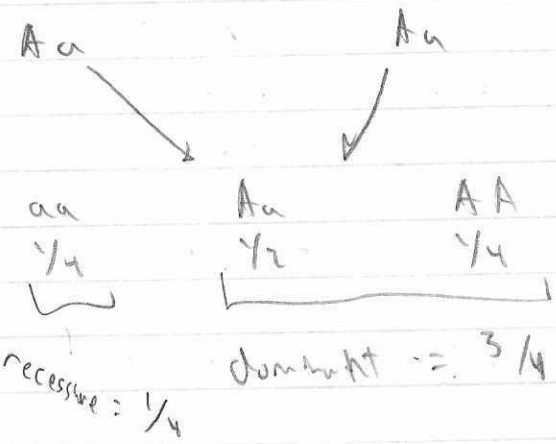
↑ But for other parameters, will not equal sample frequencies

- Parameters: numbers that summarize population
- statistic: numbers that summarize sample
- $E[X]$ = average of X over all possible samples
  = "expected value"

Mendel and Fisher

- Mendel looked at lots of phenotypes that were controlled by single loci and were dominant/recessive





- Table 1.8: Rows 1 and 2: all F2 individuals
  Rows 3 and 22: specific F2 individuals from the same cross of F1 individuals

  } seed characteristics

- Look at Row 3:

$$n_{AA} + n_{Aa} = 45, \quad n_{aa} = 12 \qquad n = 57$$

- Does 45 and 12 closely follow the $\frac{3}{4}$, $\frac{1}{4}$ proportions we expect?

chi-square

- χ^2 goodness-of-fit test

- If a category has observed count o and expected count e under some hypothesis, we can test this hypothesis w/ statistic

$$\chi^2 = \sum_{\text{categories}} \frac{(o-e)^2}{e}$$

- If hypothesis is true, o and e should be close

$\chi^2 \sim \chi^2_{df}$ where df = degrees of freedom (depends on hypothesis being tested)

↑ (#cells) − (#parameters under Null) − 1

- Expected Aa or AA = $\wedge \frac{3}{4} = 57 \cdot \frac{3}{4} = 42.75$

Expected aa = $\wedge \cdot \frac{1}{4} = 57 \cdot \frac{1}{4} = 14.25$

$$\chi^2 = \frac{(45-42.75)^2}{42.75} + \frac{(12-14.25)^2}{14.25} = 0.47$$

Compare to χ^2_1 distribution