Program of Medical Informatics
Faculty of Computers and Information
Mansoura University

Detecting Genetic Variations -II

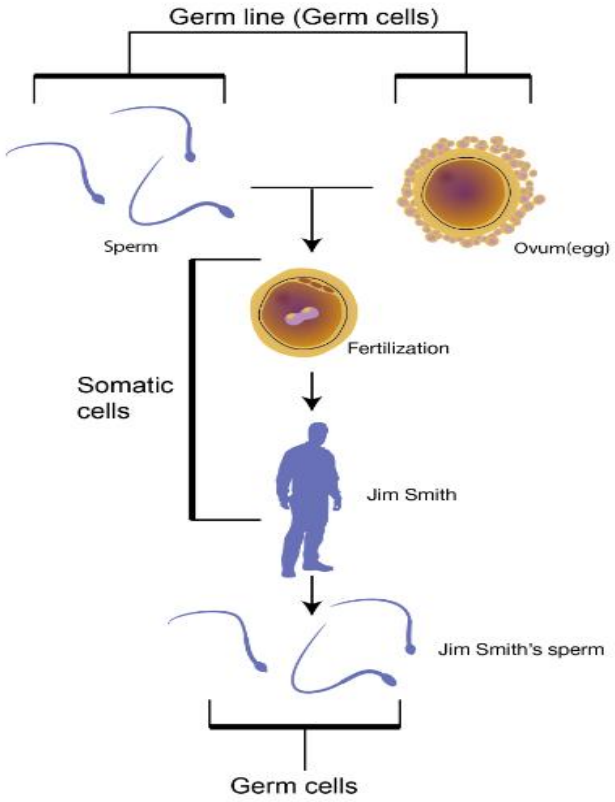
(Genomics)

Sara El-Metwally, Ph.D.
Faculty of Computers and Information,
Mansoura University, Egypt.

Email: sarah_almetwally4@mans.edu.eg
sara.elmetwally.2007@gmail.com

Office: Faculty of CIS, third floor

Notes

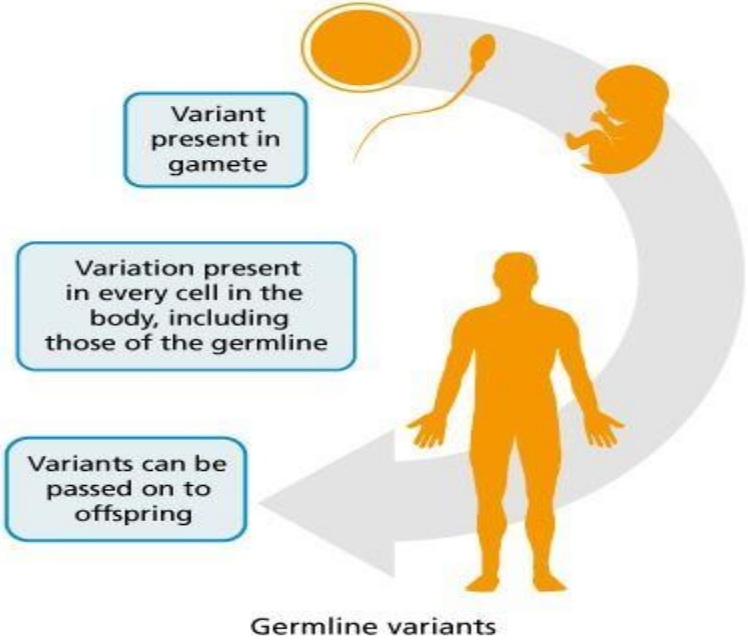


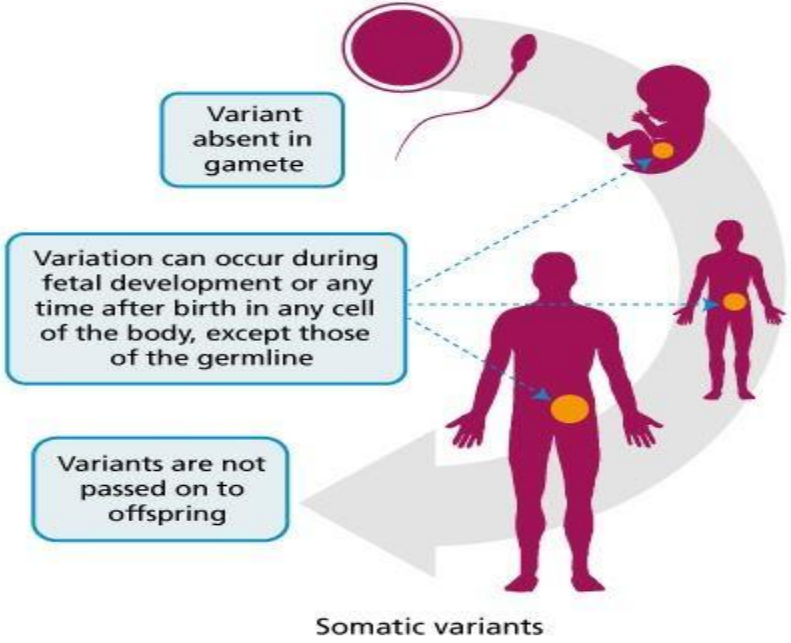
- A germ line is the sex cells (eggs and sperm) that are used by sexually reproducing organisms to pass on genes from generation to generation. Egg and sperm cells are called germ cells, in contrast to the other cells of the body that are called somatic cells.

- A gene change in a reproductive cell (egg or sperm) that becomes incorporated into the DNA of every cell in the body of the offspring. A variant (or mutation) contained within the germline can be passed from parent to offspring, and is, therefore, hereditary. Also called germline mutation.

https://www.genome.gov/genetics-glossary/germ-line

Notes





https://www.genomicseducation.hee.nhs.uk/cancer-genomics/?utm_source=Facebook&utm_medium=social&utm_campaign=SocialSignIn&utm_content=Genomics

Genetic Allele

- An allele is a variant form of a gene. Some genes have a variety of different forms, which are located at the same position, or genetic locus, on a chromosome.
- Humans are called diploid organisms because they have two alleles at each genetic locus, with one allele inherited from each parent.

https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-the-gene-pool

Genetic Allele

- Each pair of alleles represents the genotype of a specific gene.
- Genotypes are described as homozygous if there are two identical alleles at a particular locus and as heterozygous if the two alleles differ.
- Alleles contribute to the organism's phenotype, which is the outward appearance of the organism.

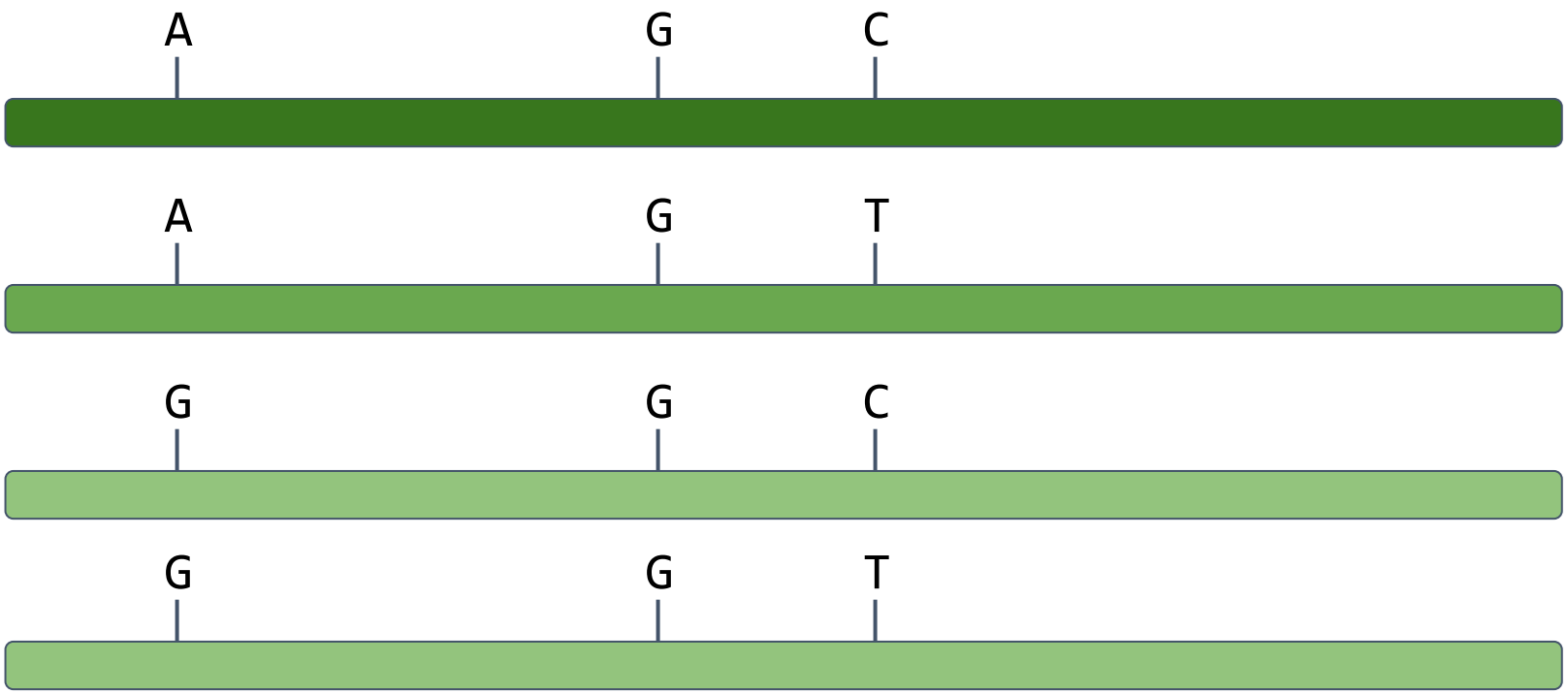
https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-the-gene-pool

Genetic Allele

- Some alleles are dominant or recessive. When an organism is heterozygous at a specific locus and carries one dominant and one recessive allele, the organism will express the dominant phenotype.
- A **haplotype** (haploid genotype) is a group of alleles in an organism that are inherited together from a single parent.
- In addition, the term "haplotype" can also refer to the inheritance of a cluster of single nucleotide polymorphisms (SNPs), which are variations at single positions in the DNA sequence among individuals.

https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-the-gene-pool

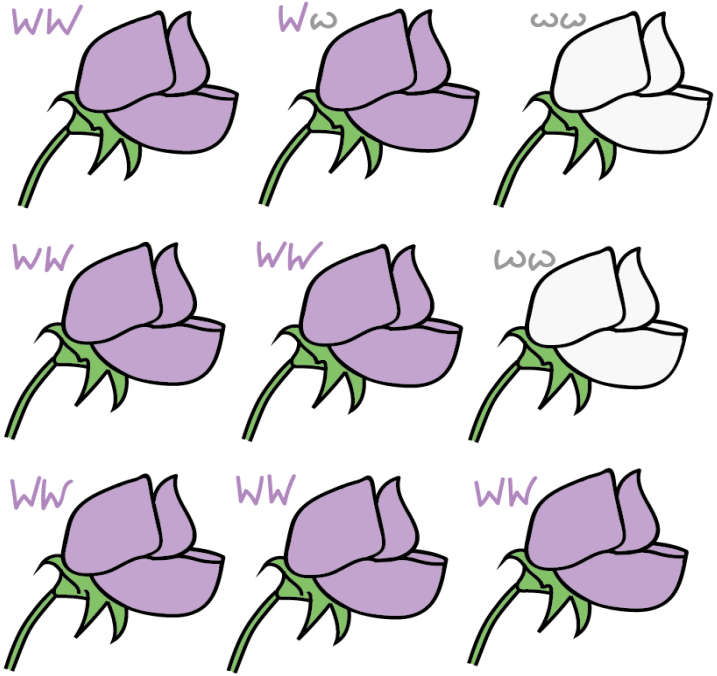
haplotype



Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Genotype vs. Allele frequency

Population of peas



GENOTYPE FREQUENCY: → How often we see each allele combo

WW, Ww, or ww

Freq. of WW = 6/9 = 0.67

Freq. of Ww = 1/9 = 0.11

Freq. of ww = 2/9 = 0.22

PHENOTYPE FREQUENCY: → How often we see white vs. purple

Freq. of purple = 7/9 = 0.78

Freq. of white = 2/9 = 0.22

ALLELE FREQUENCY: → How often we see each allele

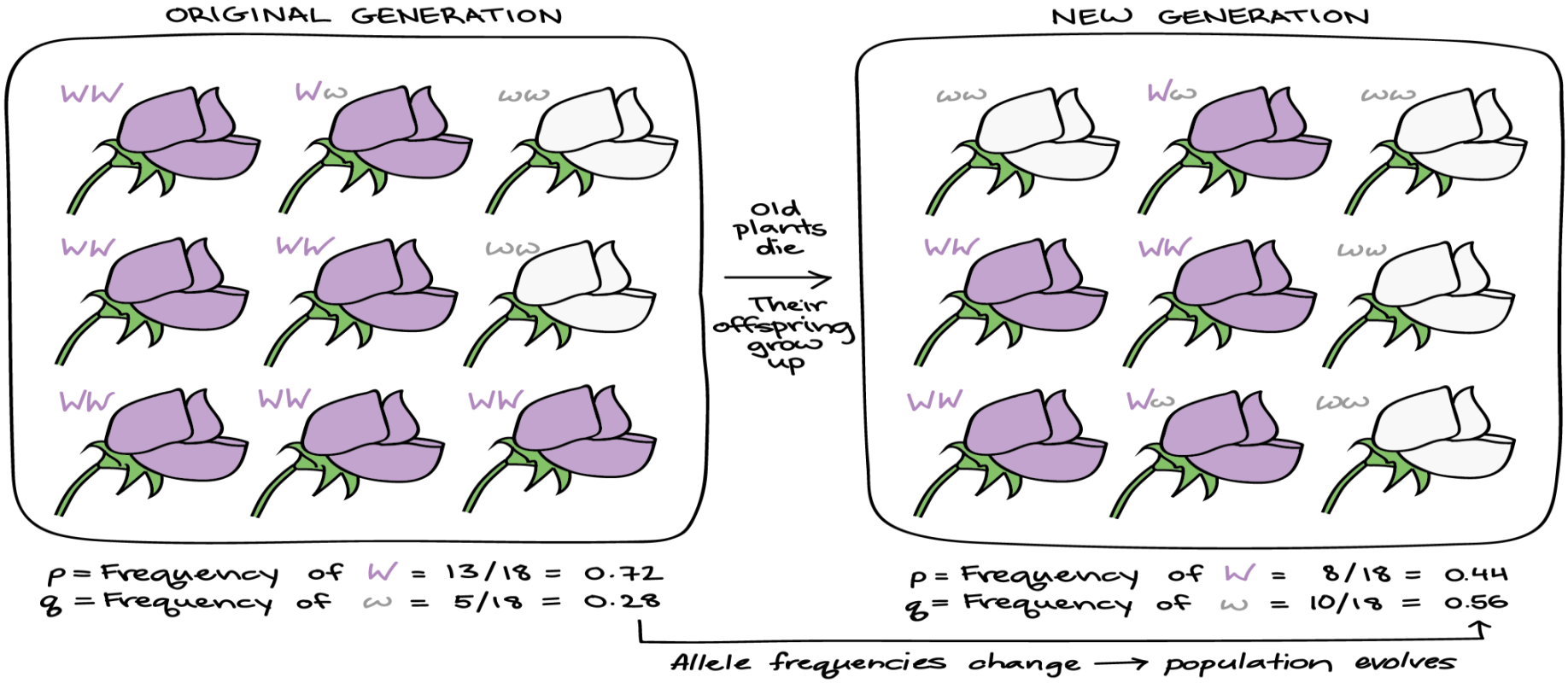
W or w

p = Freq. of W = 13/18 = 0.72

q = Freq. of w = 5/18 = 0.28

https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-the-gene-pool

Notes


ORIGINAL GENERATION
WW
Ww
ww
WW
WW
ww
WW
WW
WW
Old plants die
Their offspring grow up
NEW GENERATION
ww
Ww
ww
WW
WW
ww
WW
Ww
ww
p = Frequency of W = 13/18 = 0.72
q = Frequency of w = 5/18 = 0.28
p = Frequency of W = 8/18 = 0.44
q = Frequency of w = 10/18 = 0.56
Allele frequencies change → population evolves

Hardy-Weinberg equation

- The equation is an expression of the principle known as Hardy-Weinberg equilibrium, which states that the amount of genetic variation in a population will remain constant from one generation to the next in the absence of disturbing factors.
- To explore the Hardy-Weinberg equation, we can examine a simple genetic locus at which there are two alleles, A and a. The Hardy-Weinberg equation is expressed as:

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

https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-the-gene-pool

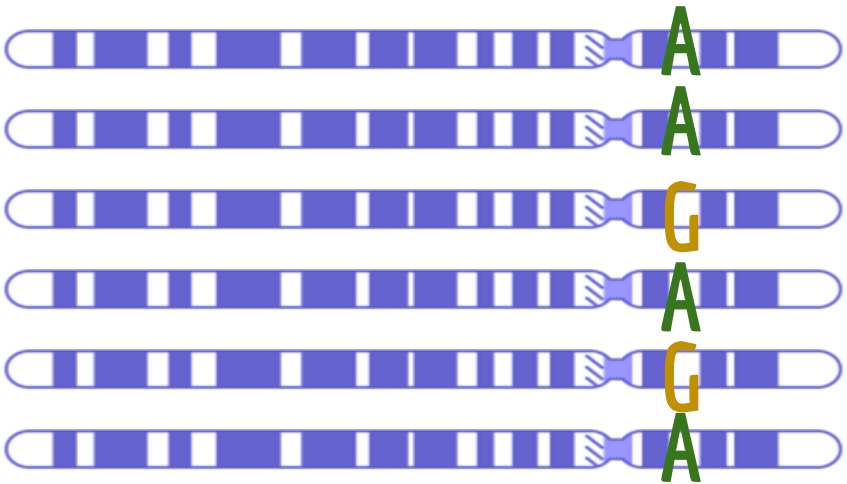
Hardy-Weinberg equation

- p is the frequency of the "A" allele and q is the frequency of the "a" allele in the population. In the equation, p^2 represents the frequency of the homozygous genotype AA, q^2 represents the frequency of the homozygous genotype aa, and 2pq represents the frequency of the heterozygous genotype Aa.
- In addition, the sum of the allele frequencies for all the alleles at the locus must be 1, so $p + q = 1$. If the p and q allele frequencies are known, then the frequencies of the three genotypes may be calculated using the Hardy-Weinberg equation.

https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-the-gene-pool

Hardy-Weinberg Equilibrium

Polymorphic loci that are biallelic (e.g., A and G alleles) have two allele frequencies, *p* and *q*.



$f(A) = p = 4/6 = 0.67$

$f(G) = q = 2/6 = 0.33$

$p + q = 1$

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Hardy-Weinberg Equilibrium

In the absence of evolutionary forces such as selection, drift, or bottlenecks, Hardy–Weinberg equilibrium states that allele and genotype frequencies in a population will remain constant from generation to generation.
If we know the allele frequencies, p and q, we can predict the genotype frequencies that should be observed (binomial expectation).

$f(A) = p = 4/6 = 0.67$
$f(G) = q = 2/6 = 0.33$

$f(AA) = p^2 = (0.67)^2 = 0.4489$
$f(AG) = 2pq = 2(0.67)(0.33) = 0.4422$
$f(GG) = q^2 = (0.33)^2 = 0.1089$

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

https://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle#Deviations_from_Hardy.E2.80.93Weinberg_equilibrium

Hardy-Weinberg Equilibrium: expected genotype freqs

$p = 0.5, q = 0.5$

$f(AA) = p^2 = (0.5)^2 = 0.25$
$f(AG) = 2pq = 2(0.5)(0.5) = 0.5$
$f(GG) = q^2 = (0.5)^2 = 0.25$

$p = 0.1, q = 0.9$

$f(AA) = p^2 = (0.1)^2 = 0.01$
$f(AG) = 2pq = 2(0.1)(0.9) = 0.18$
$f(GG) = q^2 = (0.9)^2 = 0.81$

$p = 0.01, q = 0.99$

$f(AA) = p^2 = (0.01)^2 = 0.0001$
$f(AG) = 2pq = 2(0.01)(0.99) = 0.0198$
$f(GG) = q^2 = (0.99)^2 = 0.9801$

$p = 0.001, q = 0.999$

$f(AA) = p^2 = (0.001)^2 = 0.000001$
$f(AG) = 2pq = 2(0.001)(0.999) = 0.001998$
$f(GG) = q^2 = (0.999)^2 = 0.998001$

https://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle#Deviations_from_Hardy.E2.80.93Weinberg_equilibrium

Hardy-Weinberg Equilibrium

$p = 0.1, q = 0.9$

$f(AA) = p^2 = (0.1)^2 = 0.01$
$f(AG) = 2pq = 2(0.1)(0.9) = 0.18$
$f(GG) = q^2 = (0.9)^2 = 0.81$

If we sequenced 100 individuals, how many A/G heterozygotes would we expect? How many A/A homozygotes?

https://en.wikipedia.org/wiki/Hardy%E2%80%93Weinberg_principle#Deviations_from_Hardy.E2.80.93Weinberg_equilibrium

Given genotype frequencies, calculate allele frequencies in a gene pool !

Alleles = A, a

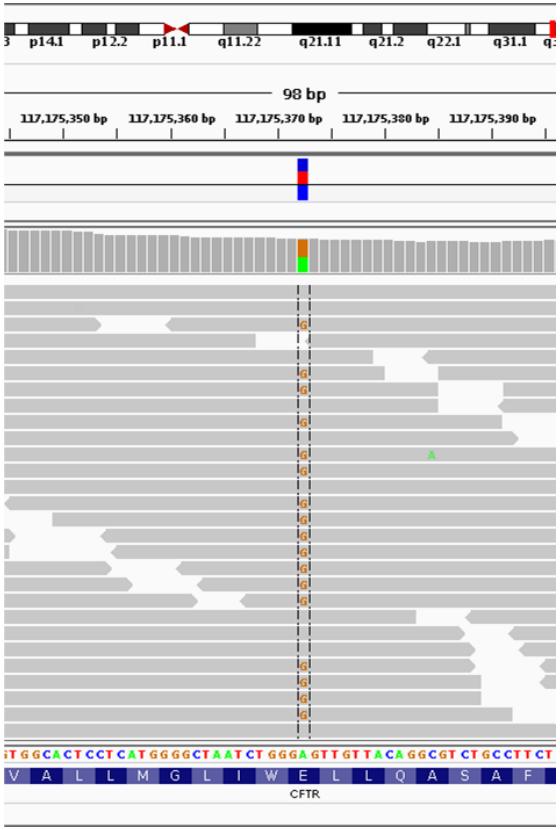
Genotypes = AA, Aa, aa

Frequency of allele A: $f(A) = f(AA) + 1/2\, f(Aa)$

Frequency of allele a: $f(a) = f(aa) + 1/2\, f(Aa)$

https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/allele-frequency-the-gene-pool

What information is needed to decide if a variant exists?



- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Mapping qualities
- Proper pairs?
- Expected polymorphism rate

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Notes

```
Mapped Reads      AATTCAGGACCCA----------------------
                  AATTCAGGACCCACACGA------------------
                  AATTCAGGACCCACACGACGGGAAGACAA-------
                  -ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
                  ----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
                  ---------ACCCACACACGACGGGTAGAGACAAGTTT
                  ---------ACCCACACACGACGGGTAGAGACAAGTTT
                  ------------GACGGGTAGAGACAAGTTTATTTTTTT
                  -------------------------TCGTTTATTTTTTT

Reference Sequence AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT
```

Notes

Potential Variant Site ?

Potential Variant Site ?

Mapped Reads

```
AATTCAGGACCCA-----------------------
AATTCAGGACCCACACGA------------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
------------GACGGGTAGAGACAAGTTTATTTTTTT
--------------------------TCGTTTATTTTTTT
```

Reference Sequence AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT

Notes

Potential Variant Site ?

Call Variant Allele?
Call Genotype?

Mapped Reads

```
AATTCAGGACCCA----------------------
AATTCAGGACCCACACGA------------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
------------GACGGGTAGAGACAAGTTTATTTTTTT
--------------------------TCGTTTATTTTTTT
```

Reference Sequence `AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT`

Ref. =A, **6 Reads**= C, **1 Reads** =A

So:

This potential site has **a Variant non reference Allele C** and **homozygous non reference genotype CC.**

Notes

Potential Variant Site ?

Call Variant Allele?
Call Genotype?

Mapped Reads

```
AATTCAGGACCCA----------------------
AATTCAGGACCCACACGA-----------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
------------GACGGGTAGAGACAAGTTTATTTTTTT
-------------------------TCGTTTATTTTTTT
```

Reference Sequence `AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT`

Ref. =A, **3 Reads**= T, **3 Reads** =A

So: This potential site has **a Variant reference Allele A and non reference Allele T** and **heterozygous genotype AT.**

Notes

Suppose that there are at maximum two alleles a and b at each site and we want to call a variants along the reference sequence.

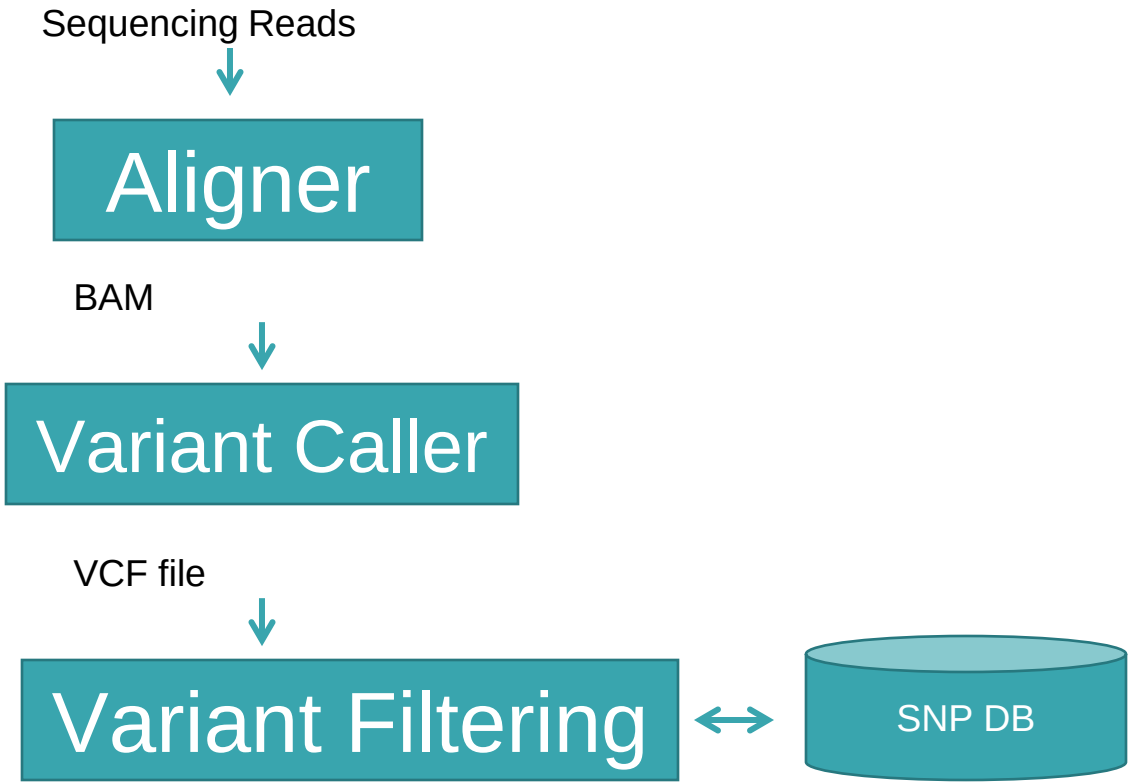
Mapped Reads

```
AATTCAGGACCCA----------------------
AATTCAGGACCCACACGA------------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
------------GACGGGTAGAGACAAGTTTATTTTTTT
-------------------------TCGTTTATTTTTTT
```

Reference Sequence `AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT`

Reference Allele	a	3	3	4	4	5			1							3								
None-Reference Allele	b	0	0	0	0	0			6							3								

Pipeline SNP detection

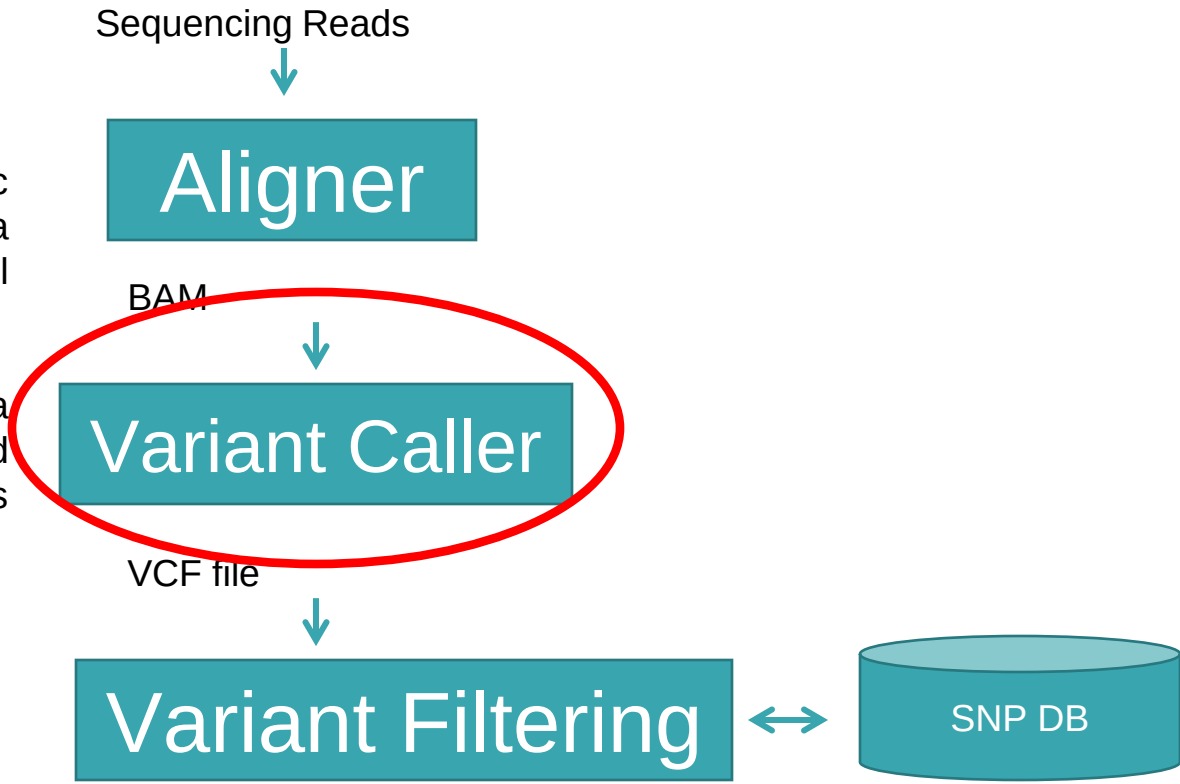


Genotyper, allele calling , are another names for variant calling

Pipeline SNP detection

- The purpose of the Probabilistic Variant Caller is to identify variants in a sample by using a probabilistic model built from read mapping data.

- This tool can detect variants in data sets from haploid (e.g. Bacteria), diploid (e.g. Human) and polyploid organisms (e.g. Cancer and higher plants).



Bayes theorem

Bayes' theorem follows from the definition of the conditional probability and relates the conditional probability $P(A|B)$ to $P(B|A)$ for two events A and B such that $P(B) \neq 0$:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- posterior
- likelihood
- prior
- normalization constant

http://www.mi.fu-berlin.de/wiki/pub/ABI/Genomics12/varcall.pdf

Bayes theorem

Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins are bent and have probability 0.6 of heads
- Type C coins are bent and have probability 0.9 of heads

Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

- Let:

  A: Event that the chosen coin was type A.

  B: Event that the chosen coin was type B.

  C: Event that the chosen coin was type C.

  D: Event that the toss was head.

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins are bent and have probability 0.6 of heads
- Type C coins are bent and have probability 0.9 of heads

Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

Prior Information before doing the experiment which is fixed and not changed if we change the experiment.
P(A)= 2/5, P(B)=2/5, P(C)=1/5.

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins are bent and have probability 0.6 of heads
- Type C coins are bent and have probability 0.9 of heads

Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

Data or Observations during doing the experiment and can be changed if the experiment is changed.

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins are bent and have probability 0.6 of heads
- Type C coins are bent and have probability 0.9 of heads

Likelihood which describes how likely that our observation or data is true given some hypotheses.
P(D|A)

Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

Example 1. There are three types of coins which have different probabilities of landing heads when tossed.

- Type A coins are fair, with probability 0.5 of heads
- Type B coins are bent and have probability 0.6 of heads
- Type C coins are bent and have probability 0.9 of heads

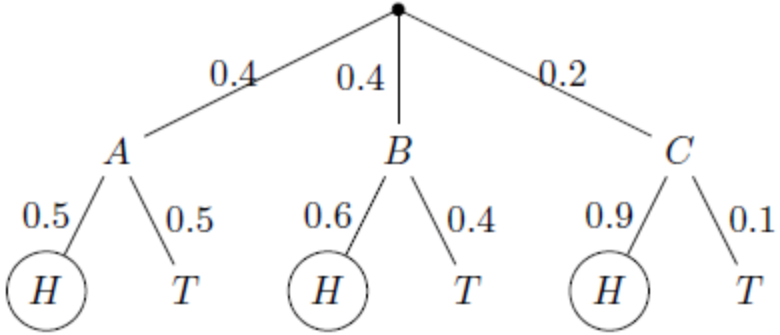
Suppose I have a drawer containing 5 coins: 2 of type A, 2 of type B, and 1 of type C. I reach into the drawer and pick a coin at random. Without showing you the coin I flip it once and get heads. What is the probability it is type A? Type B? Type C?

Posterior probability which describes what is the probability that a coin of type A, B, C given that the toss was head which our observations. P(A|D)

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

First we organize the probabilities into a tree:

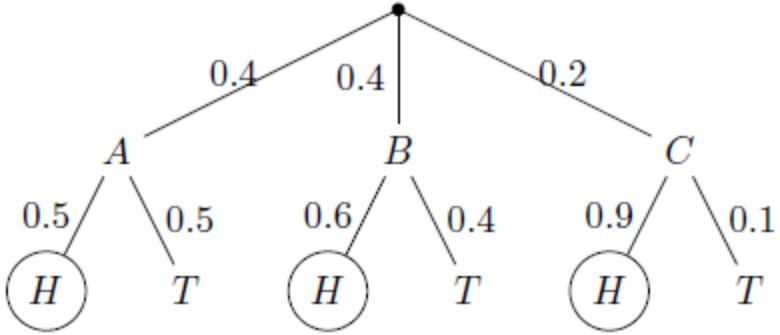


Probability tree for choosing and tossing a coin.

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

First we organize the probabilities into a tree:



Probability tree for choosing and tossing a coin.

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

Bayes' theorem says, e.g. $P(A|\mathcal{D}) = \dfrac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})}$. The denominator $P(\mathcal{D})$ is computed using the law of total probability:

$$P(\mathcal{D}) = P(\mathcal{D}|A)P(A) + P(\mathcal{D}|B)P(B) + P(\mathcal{D}|C)P(C) = 0.5 \cdot 0.4 + 0.6 \cdot 0.4 + 0.9 \cdot 0.2 = 0.62.$$

Now each of the three posterior probabilities can be computed:

$$P(A|\mathcal{D}) = \frac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})} = \frac{0.5 \cdot 0.4}{0.62} = \frac{0.2}{0.62}$$

$$P(B|\mathcal{D}) = \frac{P(\mathcal{D}|B)P(B)}{P(\mathcal{D})} = \frac{0.6 \cdot 0.4}{0.62} = \frac{0.24}{0.62}$$

$$P(C|\mathcal{D}) = \frac{P(\mathcal{D}|C)P(C)}{P(\mathcal{D})} = \frac{0.9 \cdot 0.2}{0.62} = \frac{0.18}{0.62}$$

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayes theorem

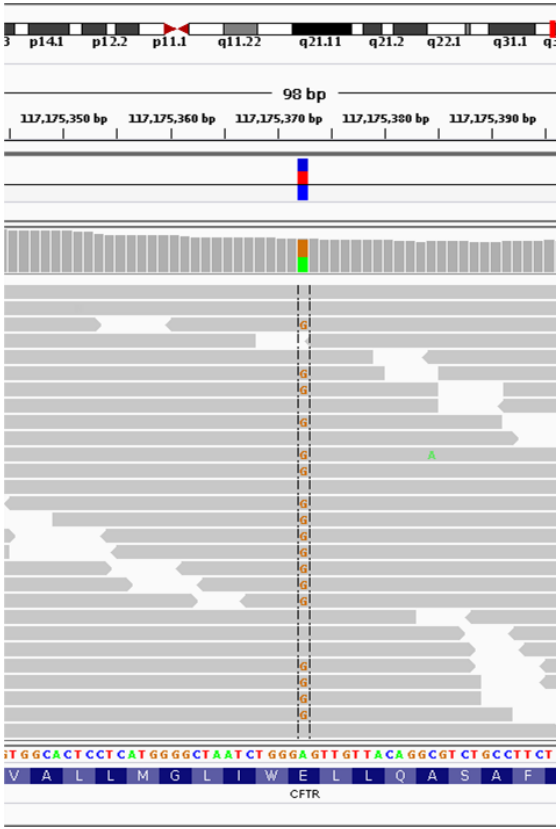
Notice that the total probability $P(\mathcal{D})$ is the same in each of the denominators and that it is the sum of the three numerators. We can organize all of this very neatly in a Bayesian update table:

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
A	0.4	0.5	0.2	0.3226
B	0.4	0.6	0.24	0.3871
C	0.2	0.9	0.18	0.2903
total	1		0.62	1

If the question is changed to decide the type of a coin, then you will compute probabilities of all types and choose the type corresponding to a maximum probability.

https://gatk.broadinstitute.org/hc/en-us/articles/360035890731-Likelihoods-and-Probabilities

Bayesian SNP calling



$$P(SNP|Data) = \frac{P(Data|SNP) * P(SNP)}{P(Data)}$$

- At each locus along the reference genome, one SNP site is called if there are a sufficient number of high-quality nucleotides to indicate a difference between the reference genome and the sample genome.

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Bayesian SNP calling

$$P(SNP|Data) = \frac{P(Data|SNP) * P(SNP)}{P(Data)}$$

- ❑ Expected Polymorphism Rate
- ❑ Expected Allele frequency at a potential site (i.e. 0.5) and then use the Hardy–Weinberg equilibrium (HWE).
- ❑ 1 polymorphic in 700 bp human, 1 in 120 for drosophila.
- ❑ Use dbSNP prior probabilities. For example, if a G/T polymorphism is reported in dbSNP, the prior probabilities are set to be 0.454 for each of the genotypes GG and TT, 0.0909 for GT and less than 10^{-4} for all other genotypes.
- ❑ If allele frequencies are known, genotype probabilities can then be calculated using the Hardy–Weinberg equilibrium (HWE) assumption or other assumptions that relate allele frequencies to genotype frequencies.
- ❑Transition Ti is more frequent than Transversion Tv.
- ❑One can assign a polymorphic rate $P_{polymorphic}$ to {AC,AG,AT,CG,CT,GT}, and (1- $P_{polymorphic}$)/4 to another non-polymorphic permutations {AA,CC,TT,GG}.

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Expected Prior Polymorphism rate

Journal List > Genome Res > v.19(6); 2009 Jun > PMC2694485

GENOME RESEARCH

CSHL Press | Journal Home | Subscriptions | eTOC Alerts | BioSupplyNet

Genome Res. 2009 Jun; 19(6): 1124–1132.
doi: 10.1101/gr.088013.108

PMCID: PMC2694485
PMID: 19420381

SNP detection for massively parallel whole-genome resequencing

Ruiqiang Li,[1,2,3] Yingrui Li,[1,3] Xiaodong Fang,[1] Huanming Yang,[1] Jian Wang,[1] Karsten Kristiansen,[1,2] and Jun Wang[1,2,4]

Author information · Article notes · Copyright and License information Disclaimer

Genome

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694485/

Expected Prior Polymorphism rate

Prior probability of genotypes

- Example: Assuming
  - heterozygous SNP rate 0.001, homozygous SNP rate 0.0005
  - Reference allele: G
  - Transition/transversion ratio 2

	A	C	G	T
A	3.33×10^{-4}	1.11×10^{-7}	6.67×10^{-4}	1.11×10^{-7}
C		8.33×10^{-5}	1.67×10^{-4}	2.78×10^{-8}
G			0.9985	1.67×10^{-4}
T				8.33×10^{-5}

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694485/

Expected Prior Polymorphism rate

Prior probability of genotypes

Other information that can be used in setting priors:

- Use dbSNP prior probability
- Use different polymorphism rate for different genomic regions
- Consider different Ti/Tv rate for exonic regions

An example of prior probability for a dbSNP G/T site used in Li et al (2009)

	A	C	G	T
A	$4.55*10^{-7}$	$9.11*10^{-8}$	$9.1*10^{-5}$	$9.1*10^{-5}$
C		$4.55*10^{-7}$	$9.1*10^{-5}$	$9.1*10^{-5}$
G			.454	.0909
T				.454

Table 35.1: Site Types for a diploid organism with example probabilities.

Site Type	Prior probability
A/A	0.2475
A/C	0.001
A/G	0.001
A/T	0.001
T/C	0.001
T/G	0.001
T/T	0.2475
G/C	0.001
C/C	0.2475
G/G	0.2475
G/-	0.001
A/-	0.001
C/-	0.001
T/-	0.001

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2694485/

http://resources.qiagenbioinformatics.com/manuals/clccancerresearchworkbench

Bayesian SNP calling

$$P(SNP|Data) = \frac{P(Data|SNP) * P(SNP)}{P(Data)}$$

- ❑ Reads that cover a variant calling position with a minimum mapping quality threshold.
- ❑ Quality score of the bases that cover the variant calling position.
- ❑ Reads count that cover a particular variant position.

Slides adapted from Aaron Quinlan: https://github.com/quinlan-lab/applied-computational-genomics

Notes

May be rare variant and not covered by the most of reads.

Mapped Reads

```
AATTCAGGACCCA----------------------
AATTCAGGACCCACACGA-----------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
-----------GACGGGTAGAGACAAGTTTATTTTTTT
-------------------------TCGTTTATTTTTTT
```

Reference Sequence `AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT`

Various Genotypes available at this locus { CC, AC, AA}

All Genotypes available at this locus { AA, AC, AG, AT, CC, CT, CG, GG, GT, TT}

Notes

b={C,C, C, A, C, C,C}, Gi={C,A}

Mapped Reads

```
AATTCAGGACCCA-----------------------
AATTCAGGACCCACACGA------------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
-----------GACGGGTAGAGACAAGTTTATTTTTTT
-------------------------TCGTTTATTTTTTT
```

Reference Sequence AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT

$$p(D|G_i) = \prod_b p(b|G_i)$$

$$p(b|G_i) = p(b|G_i^1, G_i^2) = \frac{1}{2}p(b|G_i^1) + \frac{1}{2}p(b|G_i^2)$$

https://academic.oup.com/bioinformatics/article/34/12/2012/4883351

Notes

b={C,C, C, A, C, C,C}, Gi={C,A}

Mapped Reads

```
AATTCAGGACCCA-----------------------
AATTCAGGACCCACACGA------------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
------------GACGGGTAGAGACAAGTTTATTTTTTT
-------------------------TCGTTTATTTTTTT
```

P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(A|C,A)=0.5*P(A|C)+0.5*P(A|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)

Reference Sequence AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT

$$p(D|G_i) = \prod_b p(b|G_i)$$

$$p(b|G_i) = p(b|G_i^1, G_i^2) = \frac{1}{2}p(b|G_i^1) + \frac{1}{2}p(b|G_i^2)$$

https://academic.oup.com/bioinformatics/article/34/12/2012/4883351

Notes

b={C,C, C, A, C, C,C}, Gi={C,A}

Mapped Reads

```
AATTCAGGACCCA-----------------------
AATTCAGGACCCACACGA------------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
------------GACGGGTAGAGACAAGTTTATTTTTTT
-------------------------TCGTTTATTTTTTT
```

P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(A|C,A)=0.5*P(A|C)+0.5*P(A|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)
P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)

Reference Sequence AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT

$$p(D|G_i) = \prod_b p(b|G_i)$$

$$p(b|G_i) = p(b|G_i^1, G_i^2) = \frac{1}{2}p(b|G_i^1) + \frac{1}{2}p(b|G_i^2)$$

https://academic.oup.com/bioinformatics/article/34/12/2012/4883351

Notes

b={C,C, C, A, C, C,C}, Gi={C,A}

P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)=?

Suppose Quality Score corresponding to base b = C in the read position is 30

Mapped Reads

```
AATTCAGGACCCA-----------------------
AATTCAGGACCCACACGA------------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
------------GACGGGTAGAGACAAGTTTATTTTTTT
-------------------------TCGTTTATTTTTTT
```

Reference Sequence `AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT`

$$p\left(b \middle| A\right) = \begin{cases} \frac{Q}{3} & b \neq A \\ 1 - Q & b = A \end{cases}$$

https://academic.oup.com/bioinformatics/article/34/12/2012/4883351

Notes

Suppose Quality Score corresponding to base b in the read position C is 30

Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%

Error rate

$$p\left(b \middle| A\right) = \begin{cases} \frac{Q}{3} & b \neq A \\ 1 - Q & b = A \end{cases}$$

Accuracy

P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)=?

Q=30, error rate = 0.001/3=0.00033
Q=30, accuracy=0.999

P(C|C)=0.999
P(C|A)=0.00033
Note: another model just use 0.001 for any bases that are different.

https://gatkforums.broadinstitute.org/gatk/discussion/4260/how-should-i-interpret-phred-scaled-quality-scores

Notes

Suppose Quality Score corresponding to base b in the read position C is 30

Phred Quality Score	Error	Accuracy (1 - Error)
10	1/10 = 10%	90%
20	1/100 = 1%	99%
30	1/1000 = 0.1%	99.9%
40	1/10000 = 0.01%	99.99%
50	1/100000 = 0.001%	99.999%
60	1/1000000 = 0.0001%	99.9999%

Error rate

$$p\left(b \middle| A\right) = \begin{cases} \frac{Q}{3} & b \neq A \\ 1 - Q & b = A \end{cases}$$

Accuracy

P(C|C,A)=0.5*P(C|C)+0.5*P(C|A)=?

Note: **You can incorporate different information in the model such as Mapping quality of the reads, allele frequency, polymorphic rate, reads count, etc.**

Notes

Mapped Reads

```
AATTCAGGACCCA----------------------
AATTCAGGACCCACACGA------------------
AATTCAGGACCCACACGACGGGAAGACAA-------
-ATTCAGGACAAACACGAAGGGAAGACAAGTTTATGT
----CAGGACCCACACGACGGGTAGAGACAAGTTTAT
---------ACCCACACACGACGGGTAGAGACAAGTTT
---------ACCCACACACGACGGGTAGAGACAAGTTT
------------GACGGGTAGAGACAAGTTTATTTTTTT
-------------------------TCGTTTATTTTTTT
```

Reference Sequence `AATTCAGGACCAACACGACGGGAAGATTCATGTACTTTT`

All Genotypes available at this locus { AA, AC, AG, AT, CC, CT, CG, GG, GT, TT}

P(AC| data) = .99
P(AA| data) = .00000001
P(TT| data) = .0000000001
P(CT|data) = 10^{-75}
etc

Choose genotype with highest probability

AC

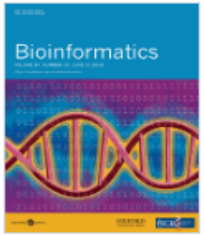
This site is a heterozygous not as we stated before as homozygous to a non reference allele

Note

OXFORD ACADEMIC

Bioinformatics

Issues | Advance articles | Submit ▾ | Purchase | Alerts | About ▾ | All Bioinformatics ▾



Volume 34, Issue 12
15 June 2018

Article Contents

Progressive approach for SNP calling and haplotype assembly using single molecular sequencing data FREE

Fei Guo, Dan Wang, Lusheng Wang

Bioinformatics, Volume 34, Issue 12, 15 June 2018, Pages 2012–2018,
https://doi.org/10.1093/bioinformatics/bty059
Published: 19 February 2018 **Article history** ▾

PDF | Split View | Cite | Permissions | Share ▾

Abstract

https://academic.oup.com/bioinformatics/article/34/12/2012/4883351

PolyBayes: the first statistically rigorous variant detection tool.

letter

A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth[1], Ian Korf[1], Mark D. Yandell[1], Raymond T. Yeh[1], Zhijie Gu[2], Hamideh Zakeri[2], Nathan O. Stitziel[1], LaDeana Hillier[1], Pui-Yan Kwok[2] & Warren R. Gish[1]

This Bayesian statistical framework has been adopted by other modern SNP/INDEL callers such as FreeBayes, GATK, and samtools

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r99

FreeBayes

Haplotype-based variant detection from short-read sequencing

Erik Garrison and Gabor Marth

July 24, 2012

Abstract

The direct detection of haplotypes from short-read DNA sequencing data requires changes to existing small-variant detection methods. Here, we develop a Bayesian statistical framework which is capable of modeling multiallelic loci in sets of individuals with non-uniform copy number. We then describe our implementation of this framework in a haplotype-based variant detector, FreeBayes.

https://arxiv.org/pdf/1207.3907.pdf
https://github.com/ekg/freebayes

GATK: Genome Analysis Toolkit

NATURE GENETICS | TECHNICAL REPORT

日本語要約

A framework for variation discovery and genotyping using next-generation DNA sequencing data

Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler & Mark J Daly

Affiliations | Contributions | Corresponding author

Nature Genetics **43**, 491–498 (2011) | doi:10.1038/ng.806

Received 27 August 2010 | Accepted 17 March 2011 | Published online 10 April 2011

http://www.nature.com/ng/journal/v43/n5/full/ng.806.html

GATK: Genome Analysis Toolkit

gatk

Best-Practices | Documentation | Blog | Forum | Download

Search

Documentation

- Getting Started
- Tool Documentation
- Methods and Algorithms
- Best Practices
- Frequently Asked Questions
- Common Problems
- Tutorials
- Dictionary
- Presentations
- Version History
- Issue Tracker

Getting Started

All you need to start using GATK today

GATK, pronounced "Gee Ay Tee Kay" (*not* "Gat-Kay"), stands for **G**enome**A**nalysis**T**ool**k**it.

It is a collection of command-line tools for analyzing high-throughput sequencing (HTS) data in formats such as SAM/BAM/CRAM and VCF, with a focus on variant discovery. The relevant file formats are defined in the hts-specs repository; see especially the SAM specification and the VCF specification.

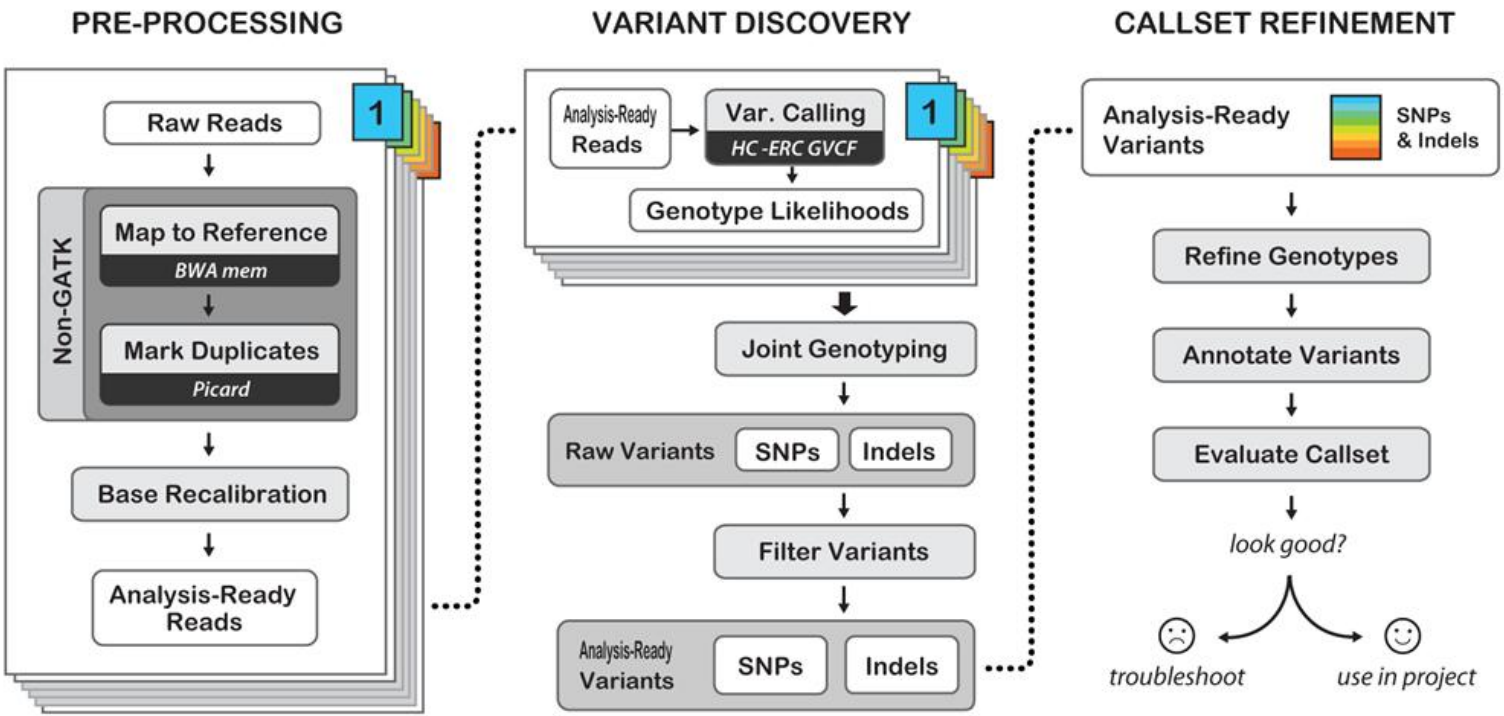
The following instructions provide the minimum requirements for getting started with GATK. Additional instructions are provided elsewhere for installing software required to run GATK Best Practices workflows and to attend a hands-on GATK workshop. For information about the complete analysis workflows we have developed for variant discovery, see the Best Practices documentation.

Download the software

The GATK command-line tools are provided as a single executable jar file. You can download a bzipped package containing the jar file from the Download page. The file name will be of the format GenomeAnalysisTK-x.y-z.tar.bz2. You will need to register for a free account on the forum and accept the licensing terms in order to access the software download.

https://software.broadinstitute.org/gatk/documentation/quickstart

GATK workflow



Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016

https://software.broadinstitute.org/gatk/img/BP workflow 3.6.png