Genetic Variation

Applied Computational Genomics, Lecture 05
https://github.com/quinlan-lab/applied-computational-genomics

Aaron Quinlan
Departments of Human Genetics and Biomedical Informatics
USTAR Center for Genetic Discovery
University of Utah
quinlanlab.org
What is genetic variation?

- Differences in DNA content or structure among individuals
  - Any two individuals have ~99.5% identical DNA.
- But the human genome is big - each haploid set of 23 chromosomes has 3.1 billion nucleotides.
  - There are >100,000,000 know genetic variants in the human genome
- Effectively infinite combinations of alleles. The details matter.

-99.8% identical DNA (differ at 1/620 - 1/750 bp

99% identical DNA

Drosophila - 1/180
<table>
<thead>
<tr>
<th>nucleotide</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>T</td>
<td>6</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
</tr>
<tr>
<td>G</td>
<td>8</td>
</tr>
<tr>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>T</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>11</td>
</tr>
<tr>
<td>G</td>
<td>12</td>
</tr>
<tr>
<td>A</td>
<td>13</td>
</tr>
<tr>
<td>T</td>
<td>14</td>
</tr>
<tr>
<td>C</td>
<td>15</td>
</tr>
<tr>
<td>G</td>
<td>16</td>
</tr>
<tr>
<td>A</td>
<td>17</td>
</tr>
<tr>
<td>T</td>
<td>18</td>
</tr>
<tr>
<td>C</td>
<td>19</td>
</tr>
<tr>
<td>G</td>
<td>20</td>
</tr>
<tr>
<td>A</td>
<td>21</td>
</tr>
<tr>
<td>T</td>
<td>22</td>
</tr>
<tr>
<td>C</td>
<td>23</td>
</tr>
<tr>
<td>G</td>
<td>24</td>
</tr>
<tr>
<td>A</td>
<td>25</td>
</tr>
<tr>
<td>T</td>
<td>26</td>
</tr>
<tr>
<td>C</td>
<td>27</td>
</tr>
<tr>
<td>G</td>
<td>28</td>
</tr>
<tr>
<td>A</td>
<td>29</td>
</tr>
<tr>
<td>T</td>
<td>30</td>
</tr>
<tr>
<td>C</td>
<td>31</td>
</tr>
<tr>
<td>G</td>
<td>32</td>
</tr>
<tr>
<td>A</td>
<td>33</td>
</tr>
<tr>
<td>T</td>
<td>34</td>
</tr>
<tr>
<td>C</td>
<td>35</td>
</tr>
<tr>
<td>G</td>
<td>36</td>
</tr>
<tr>
<td>A</td>
<td>37</td>
</tr>
<tr>
<td>T</td>
<td>38</td>
</tr>
<tr>
<td>C</td>
<td>39</td>
</tr>
<tr>
<td>G</td>
<td>40</td>
</tr>
<tr>
<td>A</td>
<td>41</td>
</tr>
<tr>
<td>T</td>
<td>42</td>
</tr>
<tr>
<td>C</td>
<td>43</td>
</tr>
<tr>
<td>G</td>
<td>44</td>
</tr>
<tr>
<td>A</td>
<td>45</td>
</tr>
</tbody>
</table>

This table represents the nucleotide sequence of a DNA or RNA strand.
Types of genetic variation

- **Single-nucleotide polymorphisms (SNPs)**
  
  "DNA spelling mistakes"

- **Insertion-deletion polymorphisms (INDELs)**
  
  "extra or missing DNA"

- **Structural variants (SVs)**
  
  "Large blocks of extra, missing or rearranged DNA"
A typical human genome

"We find that a typical [human] genome differs from the reference human genome at 4.1 million to 5.0 million sites. Although >99.9% of variants consist of SNPs and short indels, structural variants affect more bases: the typical genome contains an estimated 2,100 to 2,500 structural variants (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), affecting ~20 million bases of sequence.

Nucleotide diversity (\(\pi\)): 1/756 bp to 1/620 bp

A global reference for human genetic variation

The 1000 Genomes Project Consortium*
Why do we care?

Understanding the relationship between genetic variation and traits or disease phenotypes

Cases (have disease) vs. Controls (no disease)

Complex diseases (multiple genes contribute to risk)
Why do we care?

Genetic variants are the "bread crumbs" for tracking evolution.

http://www.fos.auckland.ac.nz/~howardross/RecreateTheResearch/RTR/images/Coalescence.png
Mutation != Polymorphism (or SNP)
It all starts with mutation

acctcccgagta  acctcccgagta
acctcccgagta  acctcccgagta
acctcccgagta  acctcccgagta
acctcccgagta  acctcccgagta
acctcccgagta  acctcccgagta
acctcccgagta  acctcccgagta

a toy population of 10 identical chromosomes
Mutation creates genetic diversity

mutation: private to this chromosome / individual
From mutation to polymorphism

acctccgagta acctcTgagta
acctccgagta acctccgagta
acctccgagta acctccgagta
acctcTgagta acctcTgagta
acctccgagta acctcTgagta
Humans are diploid.

Our genome is comprised of a paternal and a maternal "haplotype". Together, they form our "genotype"
How existing (germline) variation is inherited

Example: Mom and dad are heterozygous; that is, the zygote from which they developed was comprised of a sperm and egg with two different alleles.

Kid is homozygous

\((C/C)\)

Kid is heterozygous

\((C/T)\)

Kid is homozygous

\((T/T)\)
**de novo mutation: the birth of new variation**

Example: Mom and dad are homozygous for the same alleles.

| ♂ | ctccgag | ♀ | ctccgag |
|   | ctccgag |   | ctccgag |

New mutation occurs in father's or mother's germ cell

Note: This is a derivative chromosome of the one the father inherited from his parents. The mutation occurred in his gamete (sperm) and was passed on to the child.

♂ ctccgag → ♂ ctctgag

Kid is heterozygous owing to de novo mutation.

(C/T)
How frequently do *de novo* mutations (DNMs) occur?

Human mutation rate
- $1.1 \times 10^{-8}$/bp/generation

$\times$

Haploid genome size
- $3.1 \times 10^9$ nucleotides

= $30-40$ DNM per haploid genome,
So approximately 80 per diploid genome

(But what is the human mutation rate, really?)
Roach et al. (2010) *Science*

estimate 28 new mutations (per haploid genome) per generation via quartet sequencing
Nachman et al. (2000) *Genetics*

estimate nearly double that number via phylogenetic comparison

*Slide from Tom Sasani*
DNMs are more likely to occur in the paternal germline, and correlate with paternal age.

Slide from Tom Sasani

(a) Proportion of de novo mutations

- Paternal: $p = 1.4 \times 10^{-147}$
- Maternal: $p = 4.4 \times 10^{-6}$

(b) Number of phased de novo SNVs vs. parental age at child birth

- Paternal: $p = 4.2 \times 10^{-10}$
- Maternal: $p = 0.2$

(data from 200 ASD trios)

Yuen et al. (2016) Nature Genomic Medicine
somatic mutations

Germline mutation
- occur in sperm or egg.
- are heritable

Somatic mutation
- non-germline tissues.
- are not heritable

Somatic mutations common in cancer
vs.

compare DNA from cancer cells to healthy cells from same individual
Selection and genetic drift

What if the mutation is:

- **beneficial**
  - Positive selection
  - Negative selection

- **deleterious**
  - Negative selection

- **neutral**
  - freq. =
  - freq. ↓
  - freq. ↑

All other chromosomes:
- acctc\textcolor{red}{c}gagta
- acctctgagta

Chromosome with new allele

Genetic “drift” - random process
Genetic bottleneck

“Mixed” population
Genetic bottleneck

“Mixed” population

Bottleneck event
(migration, war, disease, nearly extinct animals
(e.g., elephant seals in Baja)
greatly reduces population size
Genetic bottleneck

“Mixed” population

Greatly reduced population diversity

Time
Out of Africa Theory
bottlenecks & reduced diversity

Nature Reviews Genetics 3, 611-621 (August 2002)
Population stratification

SNP: rs4307059
Ancestral Allele: T
Derived Allele: C
A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.
Site Frequency Spectrum - most variants are rare

Extended Data Figure 3 | Variant counts. a, The number of variants within the phase 3 sample as a function of alternative allele frequency. b, The average number of detected variants per genome with whole-sample allele frequencies <0.5% (grey bars), with the average number of singletons indicated by colours.
ExAC: exome sequencing of 60,706 humans

Analysis of protein–coding genetic variation in 60,706 humans

1 variant every 8 bases among 60,706 humans! Also, >50% of the 9 million variants discovered were present as a heterozygote in 1 out of the 60706 individuals (a "singleton")!
Allele frequency spectrum: most variants are rare

>50% of the 9 million variants discovered were present as a heterozygote in 1 out of the 60706 individuals (a "singleton"). >90% present on <=10 chromosomes sampled

http://www.nature.com/nature/journal/v536/n7616/pdf/nature19057.pdf
Each chromosome is a mosaic of alleles.

You inherit two haplotypes: one from mom, one from dad.
Meiotic recombination shuffles alleles and generates new haplotypes
Recombination frequency increases with distance between two "markers" (polymorphic sites). T.H. Morgan
One centimorgan (cM) is the equivalent to a recombination frequency of 0.01 (1%).

In humans, 1 cM corresponds to approximately 1 million bp on average.
Linkage equilibrium: random association of alleles at different loci.

A A C T G G

A G C T

G G C

G G T

25%

25%

25%

25%
Linkage disequilibrium: non-random association of alleles at different loci.

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

- 80%
- 10%
- 7%
- 3%
Therefore, knowing one allele (e.g., the first A) is a strong predictor of other alleles on a haplotype.
Sequence analysis

The variant call format and VCFtools

Petr Danecek\(^1\)*, Adam Auton\(^2\)*, Goncalo Abecasis\(^3\), Cornelis A. Albers\(^1\), Eric Banks\(^4\), Mark A. DePristo\(^4\), Robert E. Handsaker\(^4\), Gerton Lunter\(^2\), Gabor T. Marth\(^5\), Stephen T. Sherry\(^6\), Gilean McVean\(^2,7\), Richard Durbin\(^1,8\) and 1000 Genomes Project Analysis Group\(^9\)

\(^1\)Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA; \(^2\)Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK; \(^3\)Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109; \(^4\)Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141; \(^5\)Department of Biology, Boston College, MA 02467; \(^6\)National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and \(^7\)Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Address for correspondence: John Quackenbush

Associate Editor: John Quackenbush

ABSTRACT

Summary: The variant call format (VCF) is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with rich annotations. VCF is usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. The format was developed for the 1000 Genomes Project, and has also been adopted by other projects such as UK10K, dbSNP and the NHGRI Exome Project. VCFtools is a software suite that implements various utilities for processing VCF files, including validation, merging, comparing and also provides a general Perl API.

Availability: http://vcftools.sourceforge.net

Contact: rd@sanger.ac.uk

Although generic feature format (GFF) has recently been extended to standardize storage of variant information in genome variant format (GVF) (Reese et al., 2010), this is not tailored for storing information across many samples. We have designed the VCF format to be scalable so as to encompass millions of sites with genotype data and annotations from thousands of samples. We have adopted a textual encoding, with complementary indexing, to allow easy generation of the files while maintaining fast data access. In this article, we present an overview of the VCF and briefly introduce the companion VCFtools software package. A detailed format specification and the complete documentation of VCFtools are available at the VCFtools web site.
VCF format

Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFTools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">  
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">  
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">  
##FORMAT=<ID=GQ,Number=1,Type=String,Description="Genotype Quality (phred score)">  
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">  
##FORMAT=<ID=DP,Number=1,Type=String,Description="Read Depth">  
##ALT=<ID=DEL,Description="Deletion">  
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">  
##INFO=<ID=END,Number=1,Type=String,Description="End position of the variant">  
#CHROM POS ID REF ALT QUAL FILTER INFO  
1 1 ACG A,AT . PASS . .  
1 2 C T,CT . PASS . .  
1 5 A G . PASS . .  
1 100 T <DEL> . PASS . .  
```

**Reference alleles (GT=0)**

**Alternate alleles (GT>0 is an index to the ALT column)**

**Deletion**

**SNP**

**Large SV**

**Insertion**

**Other event**

**Phased data** (G and C above are on the same chromosome)
## Genotypes

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>LF1396</th>
<th>Hom. Ref.</th>
<th>Het.</th>
<th>Hom. Alt.</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr7</td>
<td>117175373</td>
<td>.</td>
<td>A</td>
<td>G</td>
<td>90</td>
<td>PASS</td>
<td>AF=0.0</td>
<td>GT</td>
<td>0/0</td>
<td>Hom. Ref.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr7</td>
<td>117175373</td>
<td>.</td>
<td>A</td>
<td>G</td>
<td>90</td>
<td>PASS</td>
<td>AF=0.5</td>
<td>GT</td>
<td>0/1</td>
<td>Het.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr7</td>
<td>117175373</td>
<td>.</td>
<td>A</td>
<td>G</td>
<td>90</td>
<td>PASS</td>
<td>AF=1.0</td>
<td>GT</td>
<td>1/1</td>
<td>Hom. Alt.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr7</td>
<td>117175373</td>
<td>.</td>
<td>A</td>
<td>G</td>
<td>0</td>
<td>PASS</td>
<td>AF=0.0</td>
<td>GT</td>
<td>./</td>
<td>Unknown</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Why would a genotype be unknown?
VCF format. A basic example

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>LF1396</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr7</td>
<td>117175373</td>
<td>.</td>
<td>A</td>
<td>G</td>
<td>90</td>
<td>PASS</td>
<td>AF=0.5</td>
<td>GT</td>
<td>0/1</td>
</tr>
</tbody>
</table>

Heterozygous A/G. The REF allele is allele "0", ALT is allele "1"
Multi-sample VCF

Heterozygous C/T.

Mom

Kid

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
<th>FORMAT</th>
<th>MOM</th>
<th>KID</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr7</td>
<td>2194169</td>
<td>.</td>
<td>C</td>
<td>T</td>
<td>210</td>
<td>PASS</td>
<td>AF=0.5</td>
<td>GT</td>
<td>0/1</td>
<td>0/1</td>
</tr>
</tbody>
</table>
Hardy-Weinberg Equilibrium

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

\[
f(A) = p = \frac{4}{6} = 0.67 \\
f(G) = q = \frac{2}{6} = 0.33
\]

\[p + q = 1\]
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).

\[
\begin{align*}
\text{f(A)} &= p = 4/6 = 0.67 \\
\text{f(G)} &= q = 2/6 = 0.33 \\
\text{f(AA)} &= p^2 = (0.67)^2 = 0.4489 \\
\text{f(AG)} &= 2pq = 2(0.67)(0.33) = 0.4422 \\
\text{f(GG)} &= q^2 = (0.33)^2 = 0.1089 \\
\end{align*}
\]

\[p^2 + 2pq + q^2 = 1\]
Hardy-Weinberg Equilibrium: expected genotype freqs

\[
p = 0.5, \quad q = 0.5
\]
\[
f(AA) = p^2 = (0.5)^2 = 0.25
\]
\[
f(AG) = 2pq = 2(0.5)(0.5) = 0.5
\]
\[
f(GG) = q^2 = (0.5)^2 = 0.25
\]

\[
p = 0.1, \quad q = 0.9
\]
\[
f(AA) = p^2 = (0.1)^2 = 0.01
\]
\[
f(AG) = 2pq = 2(0.1)(0.9) = 0.18
\]
\[
f(GG) = q^2 = (0.9)^2 = 0.81
\]

\[
p = 0.01, \quad q = 0.99
\]
\[
f(AA) = p^2 = (0.01)^2 = 0.0001
\]
\[
f(AG) = 2pq = 2(0.01)(0.99) = 0.0198
\]
\[
f(GG) = q^2 = (0.99)^2 = 0.9801
\]

\[
p = 0.001, \quad q = 0.999
\]
\[
f(AA) = p^2 = (0.001)^2 = 0.000001
\]
\[
f(AG) = 2pq = 2(0.001)(0.999) = 0.001998
\]
\[
f(GG) = q^2 = (0.999)^2 = 0.998001
\]
Hardy-Weinberg Equilibrium

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

If we sequenced 100 individuals, how many A/A homozygotes?

\[ p = 0.1, \quad q = 0.9 \]

\[ f(AA) = p^2 = (0.1)^2 = 0.01 \]
\[ f(AG) = 2pq = 2(0.1)(0.9) = 0.18 \]
\[ f(GG) = q^2 = (0.9)^2 = 0.81 \]
Hardy-Weinberg Equilibrium
Hardy-Weinberg Equilibrium - example

The frequency of allele "Z" at a given locus on chromosome 7 is 0.3. What proportion of individuals do we expect to be heterozygous for the Z and Q alleles?
Deviations from Hardy-Weinberg Equilibrium

- Inbreeding
- Population bottlenecks
- Positive selection
- Purifying selection
- Random genetic drift

Example: a recessive, disease causing allele.
Expect $p^2$ homozygotes for the recessive allele, yet observe significantly less than $p^2 \times$ the number of individuals tested

http://www.nature.com/scitable/knowledge/library/the-hardy-weinberg-principle-13235724