Assignment 9

Mayank Choudhary
Bio5488
25th March 2016

Adapted with permission from http://biocomicals.blogspot.com

Assignment 9: Profile Genetic Variation

Goal:
Given VCF files, profile various classes of genetic variation and study basic principles of genetic.

Input:
VCF file containing

- SNV and indels: `SNV_indel.biallelic.vcf`
- SVs: `sv.reclassed.filtered.vcf`

Output:
Basic VCF parser!! Mendelian violations!!

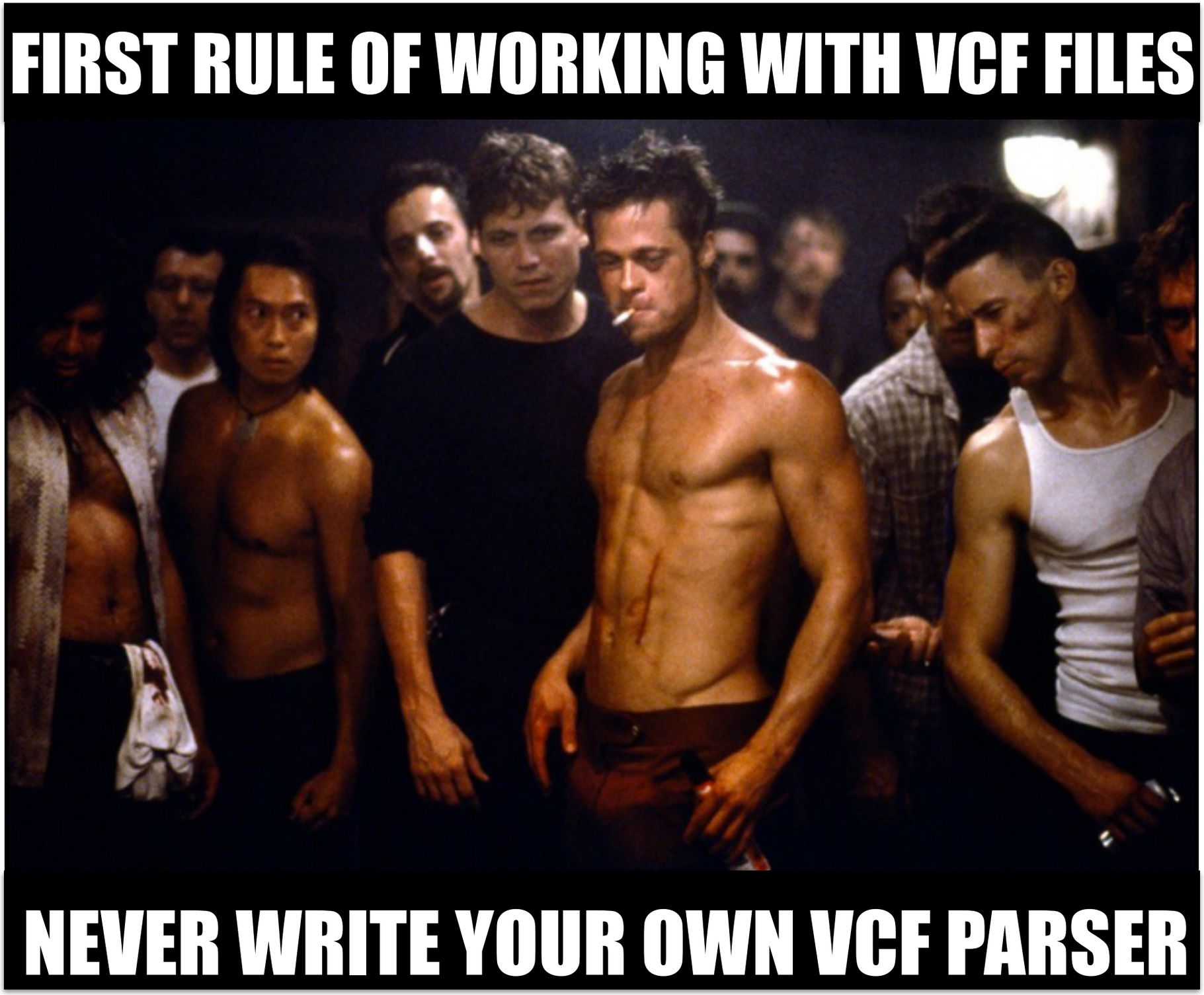
*

**if you survive to tell the tale*

CAUTION! **DO NOT** copy to `/work/`

memegenerator.net

First rule first



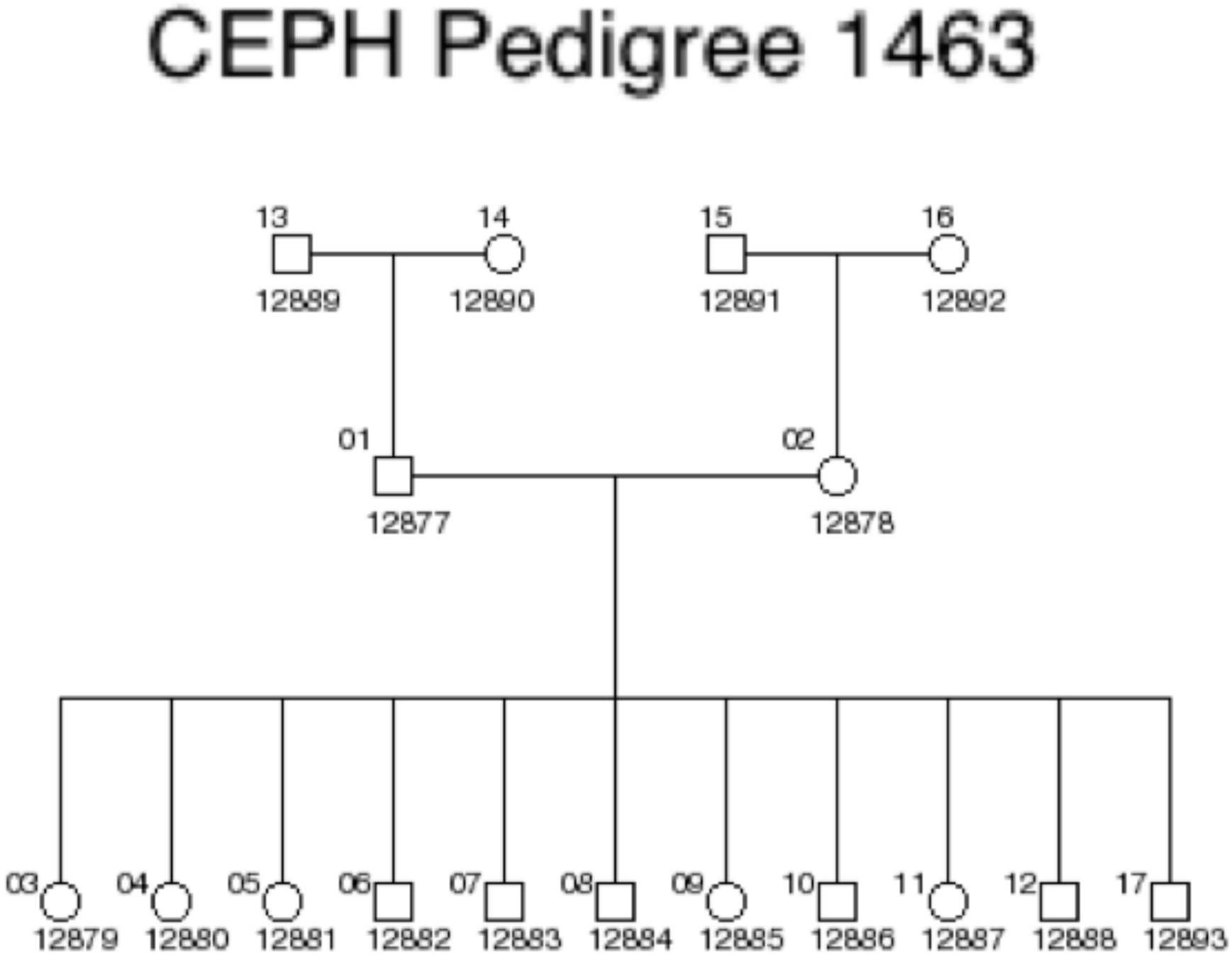
memegenerator.net

~~First rule first~~ Rules are meant to be broken!

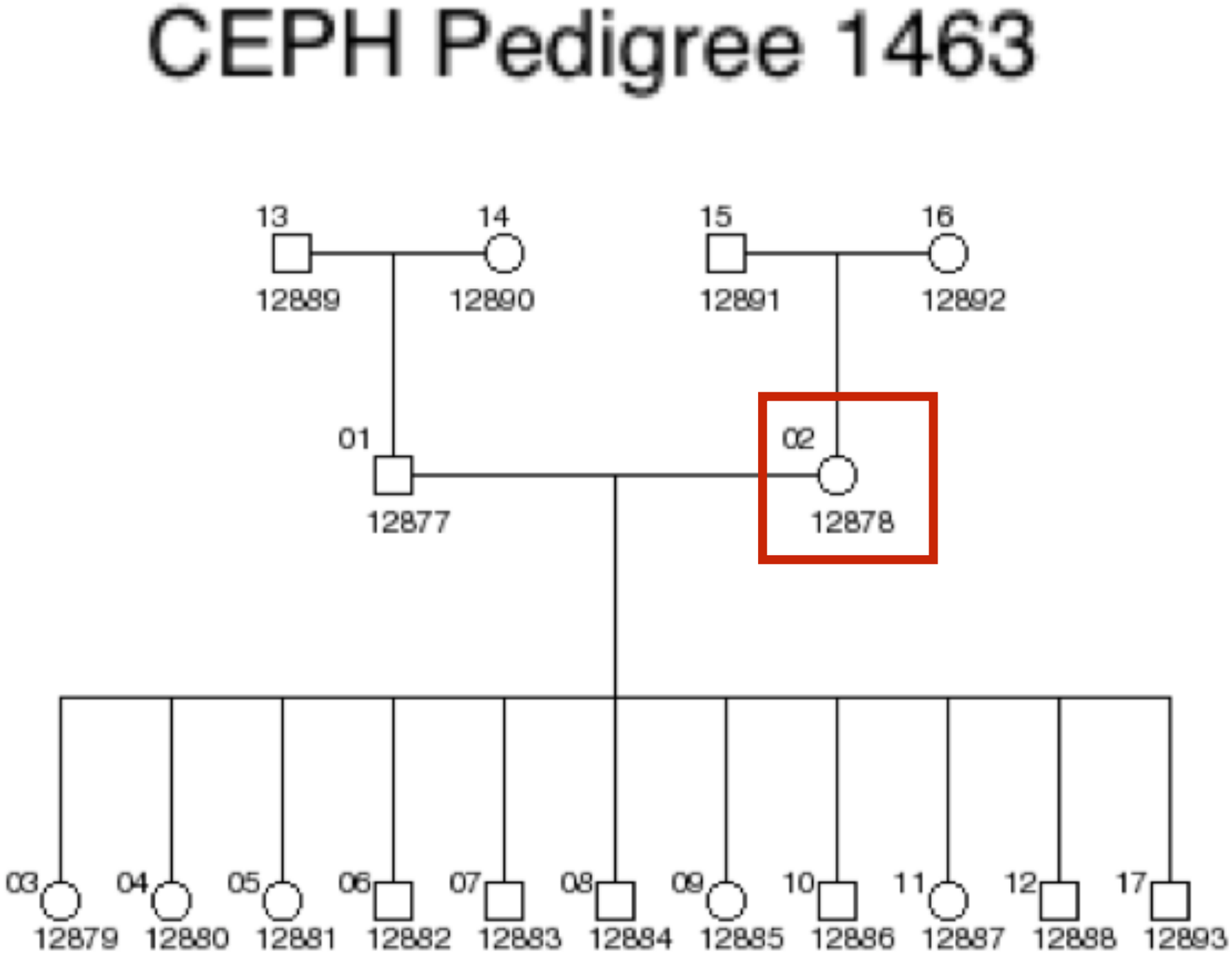
memegenerator.net

Variant calls from a 17-member family

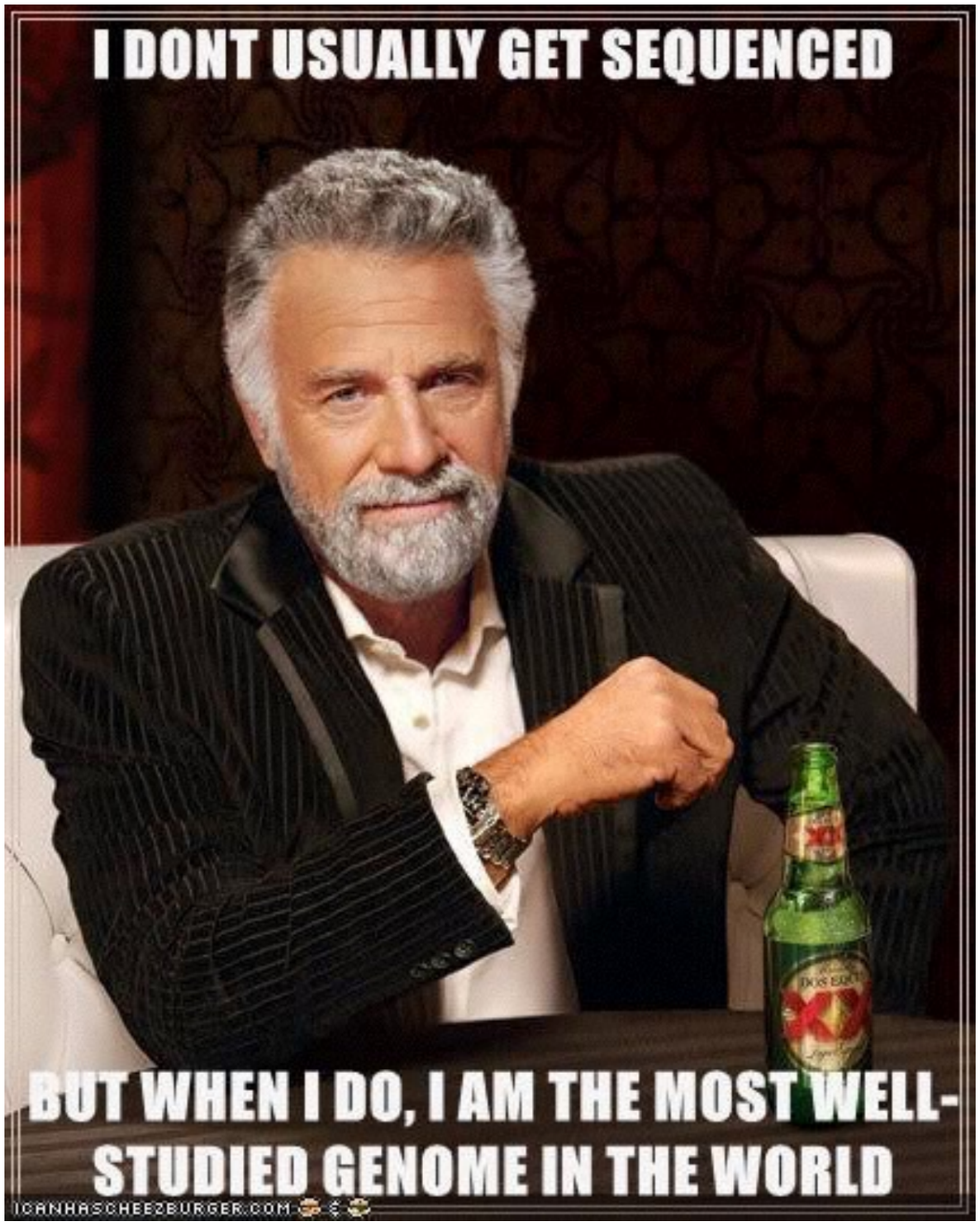
Extremely (re)productive F1s!



Extremely (re)productive F1s!

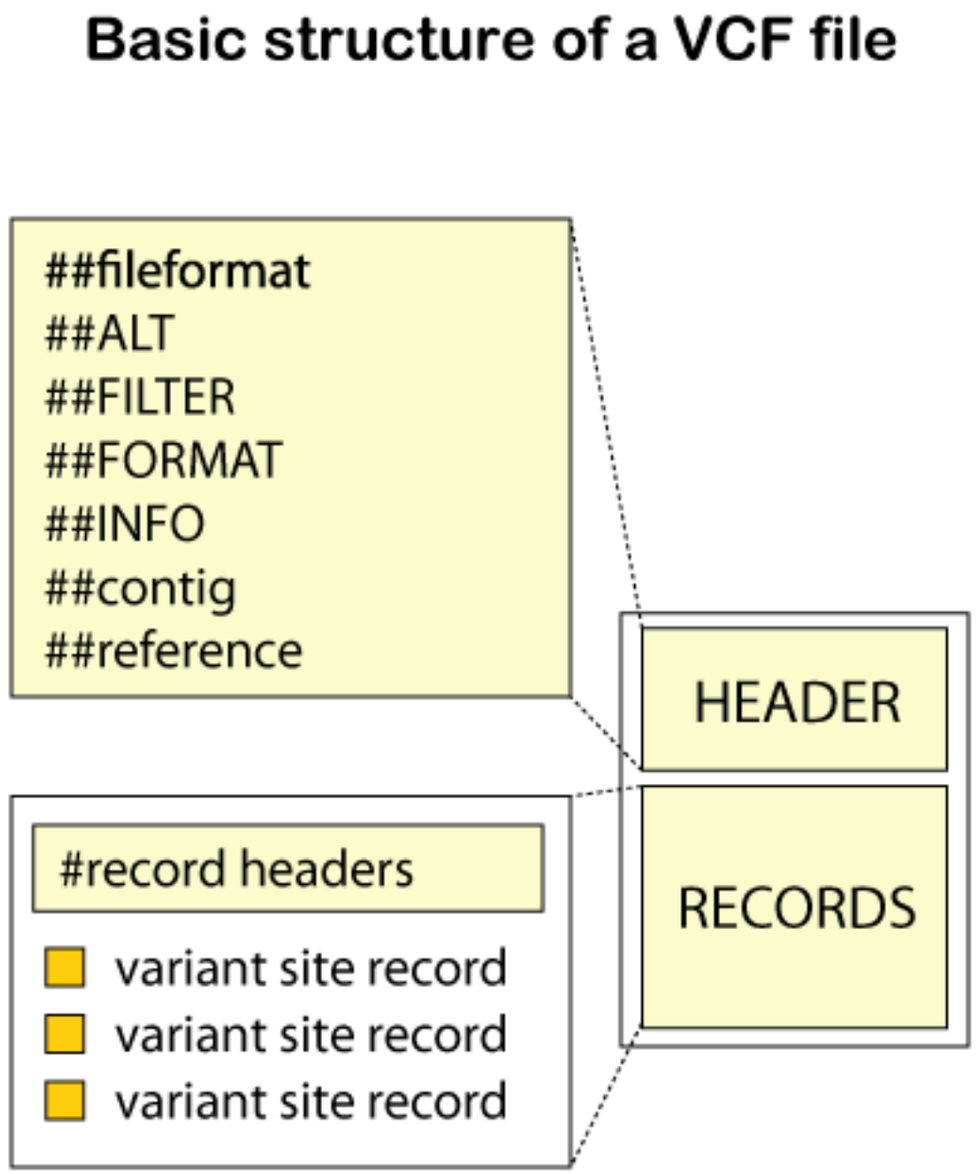


NA12878–the Jonathan Goldsmith of genomes!



memegenerator.net

Anatomy of a VCF file



http://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it

Header of a VCF file

```
##FORMAT=<ID=AS,Number=A,Type=Integer,Description="Alternate allele split-read observation count, with partial observations recorded fractionally">
##FORMAT=<ID=RP,Number=1,Type=Integer,Description="Reference allele paired-end observation count, with partial observations recorded fractionally">
##FORMAT=<ID=AP,Number=A,Type=Integer,Description="Alternate allele paired-end observation count, with partial observations recorded fractionally">
##FORMAT=<ID=AB,Number=A,Type=Float,Description="Allele balance, fraction of observations from alternate allele, QA/(QR+QA)">
##FORMAT=<ID=CN,Number=1,Type=Float,Description="Copy number of structural variant segment.">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA12877 NA12878 NA12879 NA12880 NA12881 NA12882 NA12882 NA12883 NA12884 NA12885 NA12886 NA12887 NA12888 NA12889 NA12890 NA12891 NA12892 NA12893
[mayank@genomic:/home/assignments/assignment9]$ c

[mayank@genomic:/home/assignments/assignment9]$ head -n53 sv.reclassed.filtered.vcf
##fileformat=VCFv4.2
##fileDate=20160313
##reference=
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=POS,Number=1,Type=Integer,Description="Position of the variant described in this record">
##INFO=<ID=SVLEN,Number=.,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=STRANDS,Number=.,Type=String,Description="Strand orientation of the adjacency in BEDPE format (DEL:+-, DUP:-+, INV:++/--)">
##INFO=<ID=IMPRECISE,Number=0,Type=Flag,Description="Imprecise structural variation">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS95,Number=2,Type=Integer,Description="Confidence interval (95%) around POS for imprecise variants">
##INFO=<ID=CIEND95,Number=2,Type=Integer,Description="Confidence interval (95%) around END for imprecise variants">
##INFO=<ID=MATEID,Number=.,Type=String,Description="ID of mate breakends">
##INFO=<ID=EVENT,Number=1,Type=String,Description="ID of event associated to breakend">
##INFO=<ID=SECONDARY,Number=0,Type=Flag,Description="Secondary breakend in a multi-line variants">
##INFO=<ID=SU,Number=.,Type=Integer,Description="Number of pieces of evidence supporting the variant across all samples">
##INFO=<ID=PE,Number=.,Type=Integer,Description="Number of paired-end reads supporting the variant across all samples">
##INFO=<ID=SR,Number=.,Type=Integer,Description="Number of split reads supporting the variant across all samples">
##INFO=<ID=EV,Number=.,Type=String,Description="Type of LUMPY evidence contributing to the variant call">
##INFO=<ID=PRPOS,Number=.,Type=String,Description="LUMPY probability curve of the POS breakend">
##INFO=<ID=PREND,Number=.,Type=String,Description="LUMPY probability curve of the END breakend">
##INFO=<ID=SNAME,Number=.,Type=String,Description="Source sample name">
##INFO=<ID=ALG,Number=1,Type=String,Description="Evidence PDF aggregation algorithm">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=NSAMP,Number=1,Type=Integer,Description="Number of samples with non-reference genotypes">
##INFO=<ID=MSQ,Number=1,Type=Float,Description="Mean sample quality of positively genotyped samples">
##ALT=<ID=DEL,Description="Deletion">
##ALT=<ID=DUP,Description="Duplication">
##ALT=<ID=INV,Description="Inversion">
##ALT=<ID=DUP:TANDEM,Description="Tandem duplication">
##ALT=<ID=INS,Description="Insertion of novel sequence">
##ALT=<ID=CNV,Description="Copy number variable region">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=SU,Number=1,Type=Integer,Description="Number of pieces of evidence supporting the variant">
##FORMAT=<ID=PE,Number=1,Type=Integer,Description="Number of paired-end reads supporting the variant">
##FORMAT=<ID=SR,Number=1,Type=Integer,Description="Number of split reads supporting the variant">
##FORMAT=<ID=BD,Number=1,Type=Integer,Description="Amount of BED evidence supporting the variant">
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT=<ID=SQ,Number=1,Type=Float,Description="Phred-scaled probability that this site is variant (non-reference in this sample">
##FORMAT=<ID=GL,Number=G,Type=Float,Description="Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for each possible genotype generated from the reference and alternate alleles given the sample ploidy">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count, with partial observations recorded fractionally">
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observations, with partial observations recorded fractionally">
##FORMAT=<ID=QR,Number=1,Type=Integer,Description="Sum of quality of reference observations">
##FORMAT=<ID=QA,Number=A,Type=Integer,Description="Sum of quality of alternate observations">
##FORMAT=<ID=RS,Number=1,Type=Integer,Description="Reference allele split-read observation count, with partial observations recorded fractionally">
##FORMAT=<ID=AS,Number=A,Type=Integer,Description="Alternate allele split-read observation count, with partial observations recorded fractionally">
##FORMAT=<ID=RP,Number=1,Type=Integer,Description="Reference allele paired-end observation count, with partial observations recorded fractionally">
##FORMAT=<ID=AP,Number=A,Type=Integer,Description="Alternate allele paired-end observation count, with partial observations recorded fractionally">
##FORMAT=<ID=AB,Number=A,Type=Float,Description="Allele balance, fraction of observations from alternate allele, QA/(QR+QA)">
##FORMAT=<ID=CN,Number=1,Type=Float,Description="Copy number of structural variant segment.">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA12877 NA12878 NA12879 NA12880 NA12881 NA12882 NA12882 NA12883 NA12884 NA12885 NA12886 NA12887 NA12888 NA12889 NA12890 NA12891 NA12892 NA12893
[mayank@genomic:/home/assignments/assignment9]$
```

http://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it

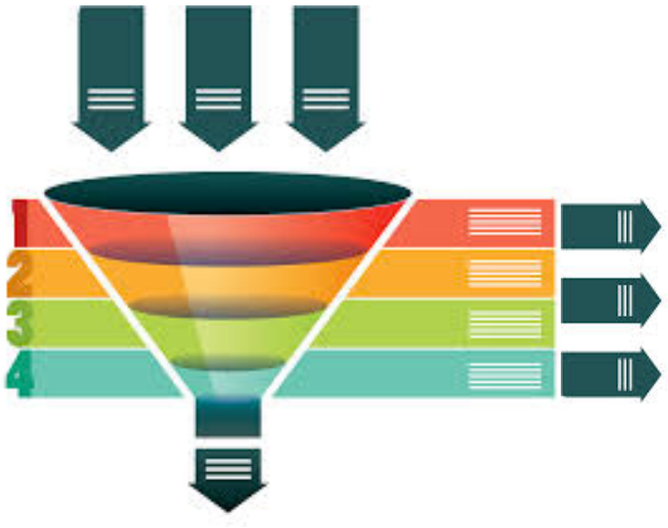
Records in a VCF file

DEL

SNV

INS

DEL

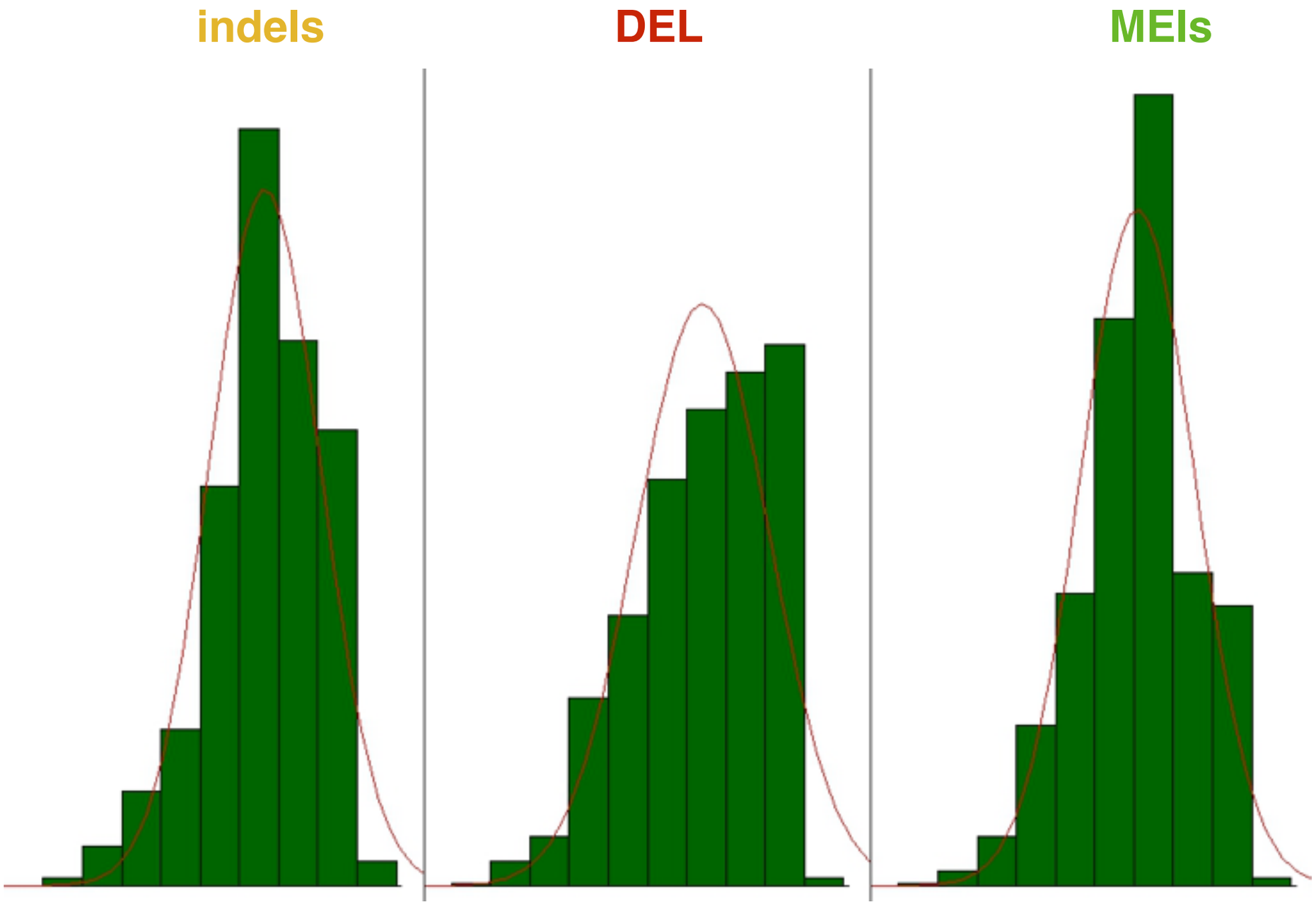


Total GV

Profiling counts

Class of genome variation	count
SNVs	…….
indels	……
DEL	….
DUP	…
INV	..
MEIs	…
BNDs	….

Plot the size distributions



Zygosity explained

GT field

./.

Reference Genome

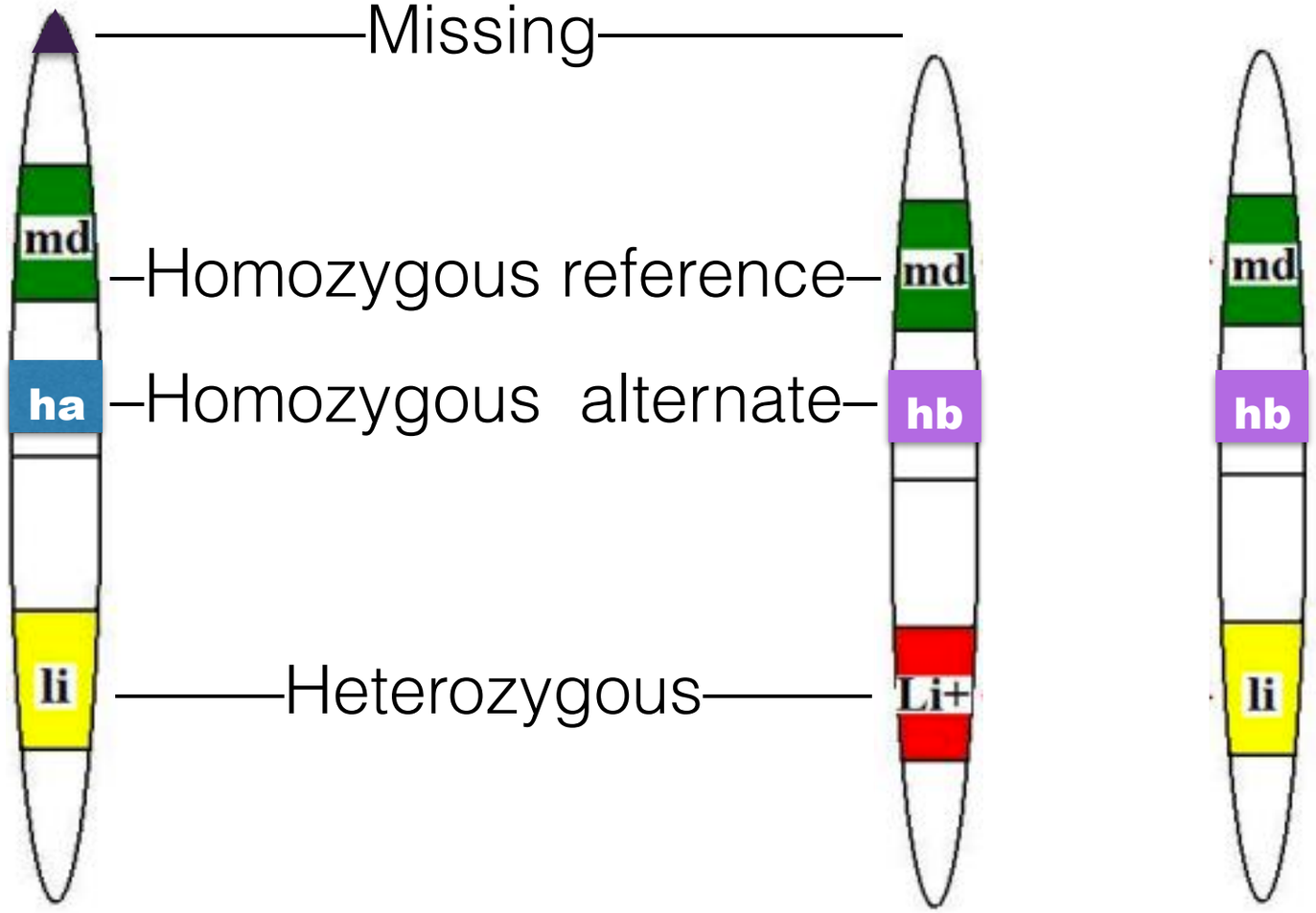
One pair of homologous chromosomes

Missing

0/0 –Homozygous reference–

1/1 –Homozygous alternate–

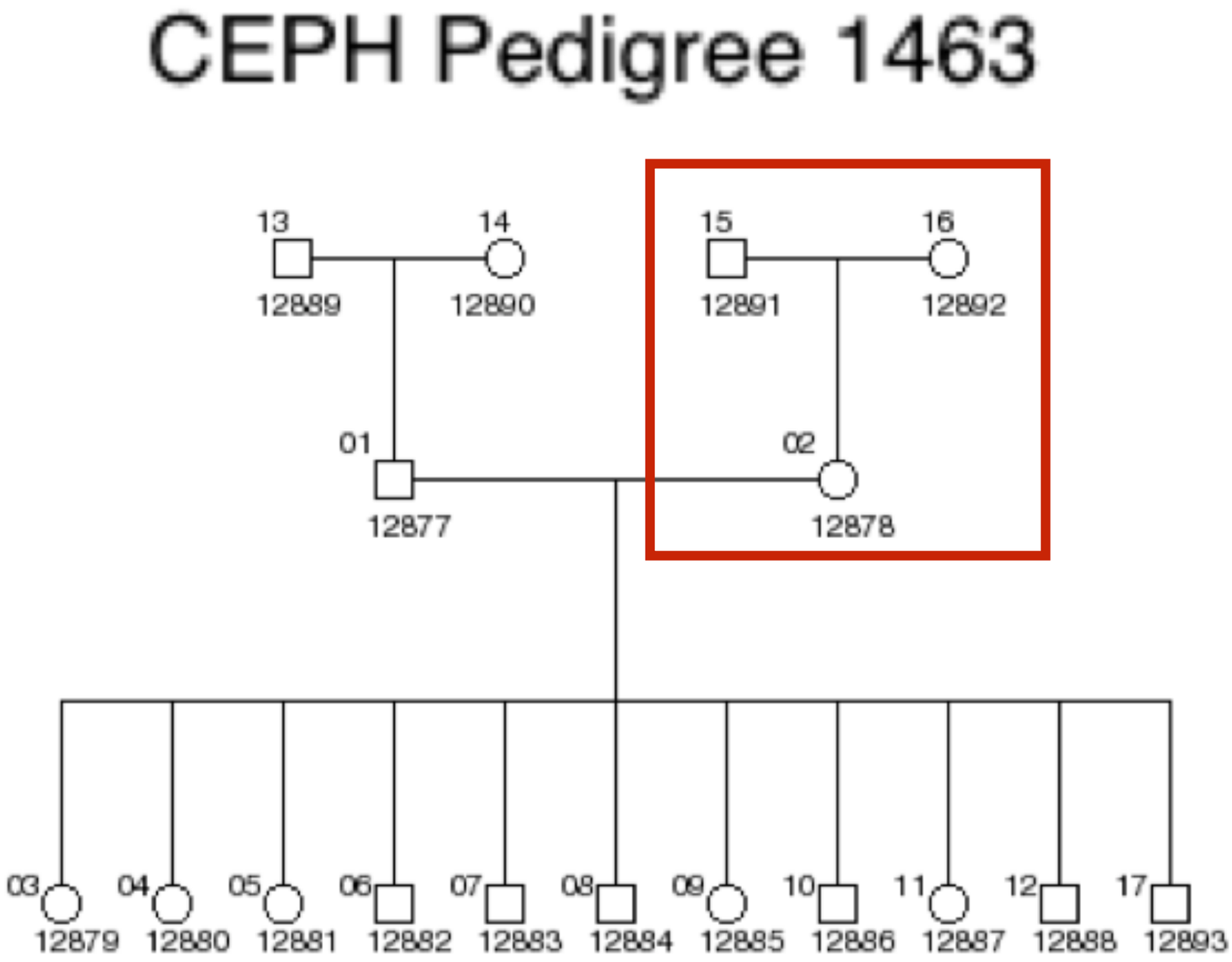
0/1 ———Heterozygous———



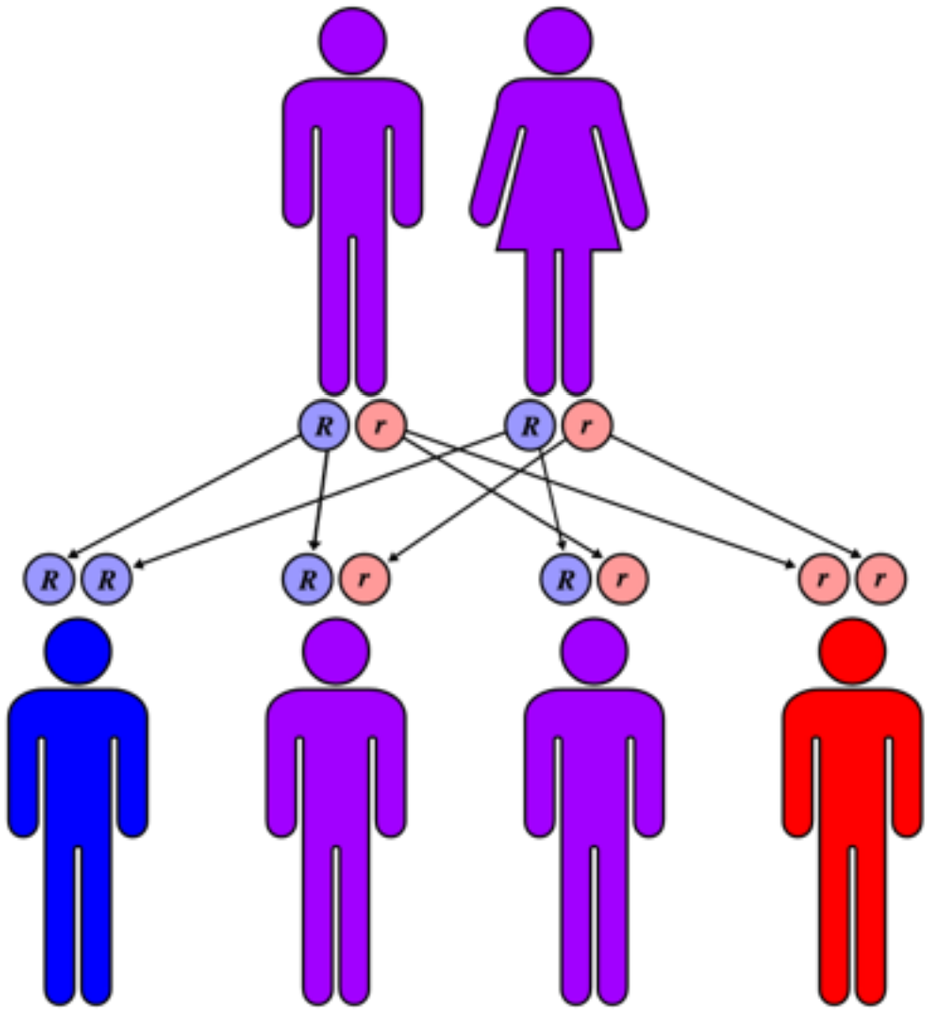
A wise (now dead) man once said…

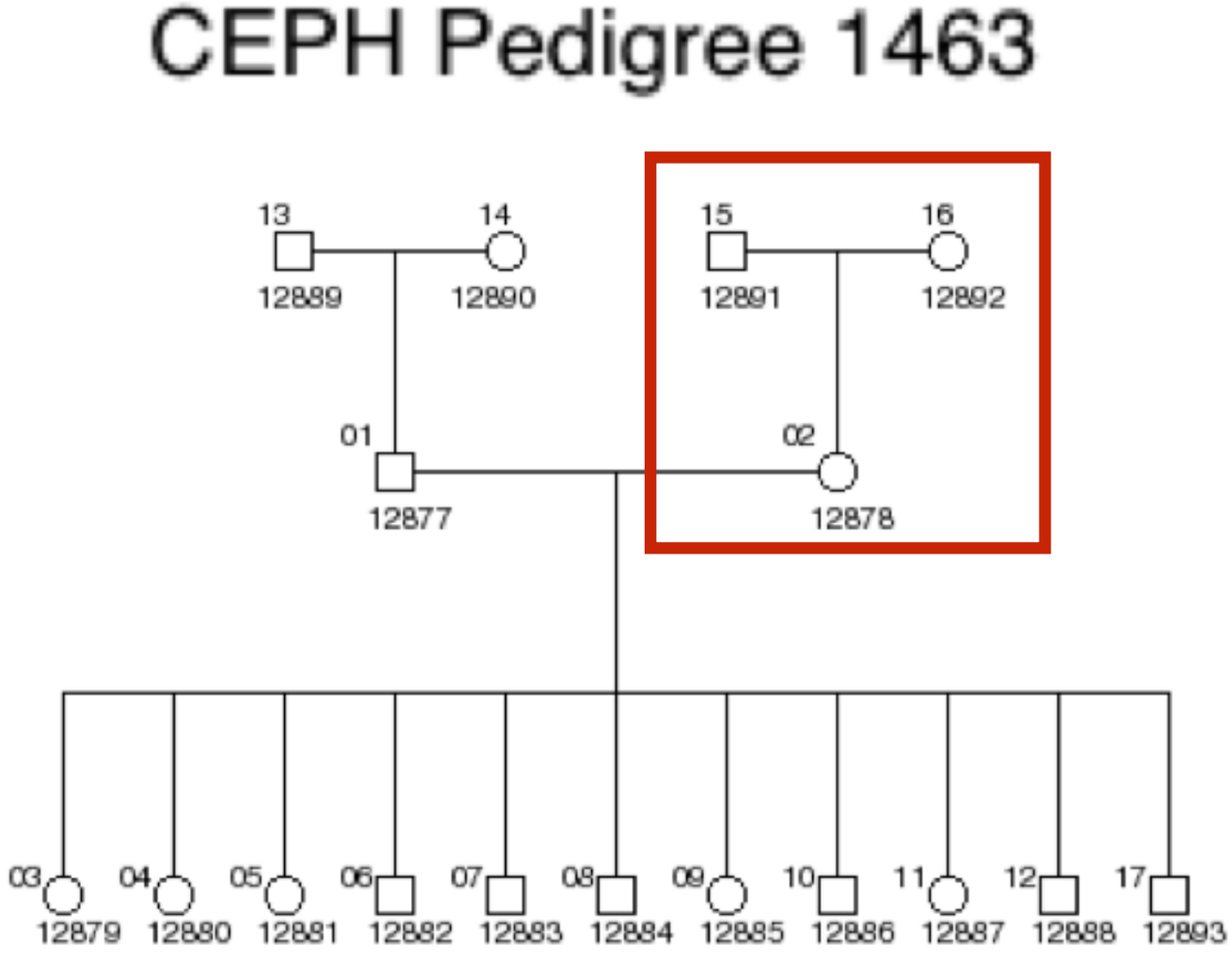
memegenerator.net

Trio analysis to look for violations of Mendel's Law of Segregation



Trio analysis to look for violations of Mendel's Law of Segregation



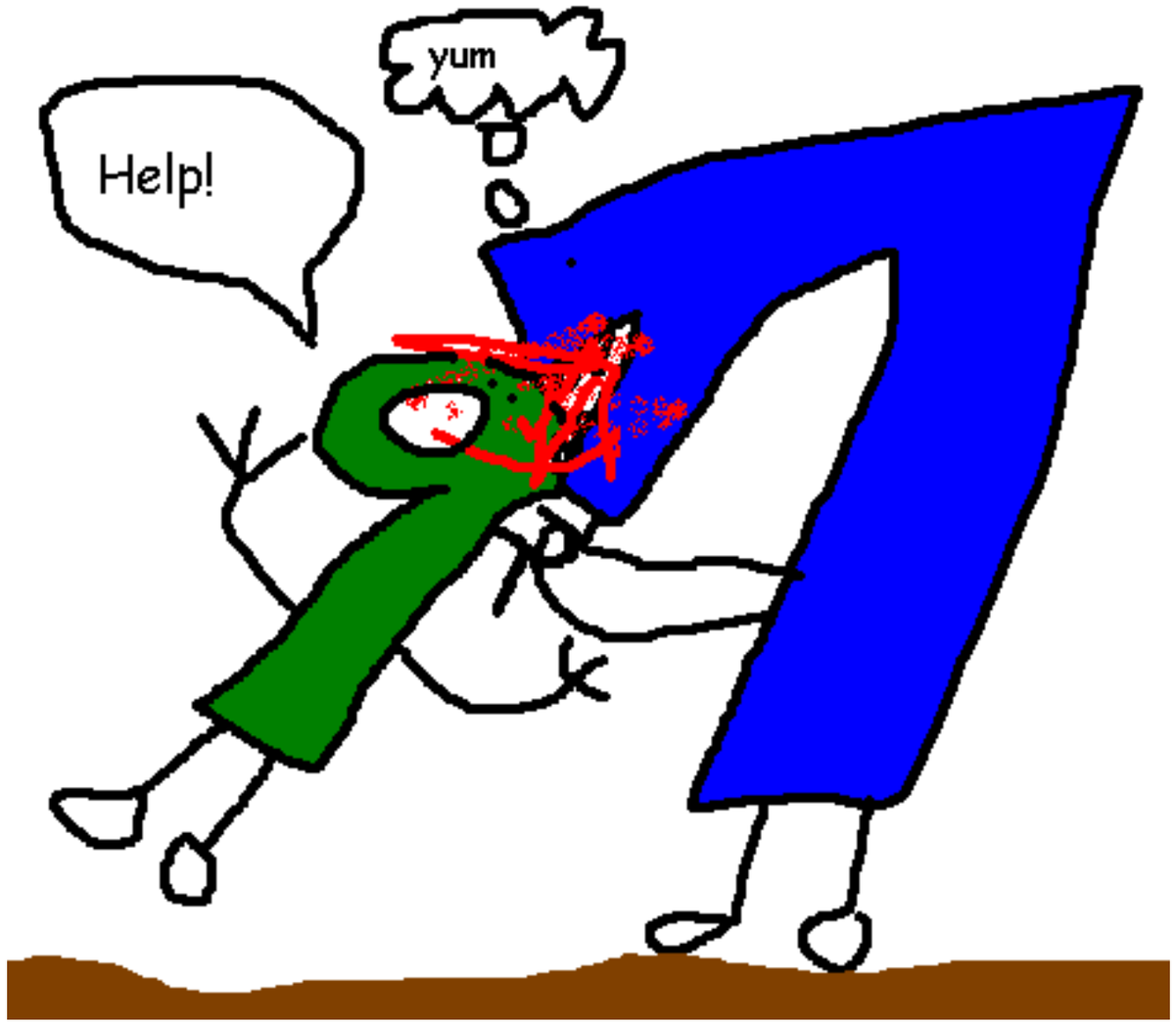


http://commons.wikimedia.org/wiki/File:Autorecessive.svg

WHY?

memegenerator.net

BUT WHERE'S QUESTION 9?



http://en.illogicopedia.org/wiki/Why_Did_Seven_Eat_Nine%3F

GQ—Genotype Quality

The following formula relates a given GQ value X to the probability that the genotype call is INCORRECT:

$X = -10*\log_{10}(\text{Probability(genotype call is incorrect)})$,

or $\text{Probability(genotype call is incorrect)} = 10^{-X/10}$

For instance, a GQ value of 20 means that you are 99% sure your genotype call is correct, or there is a 1% chance your genotype call is incorrect.

http://riddhitubes.com/images/quality-stamp.png

Assignment 9 requirements

- Input files located in `/home/assignments/assignment9/`
- Important: **DO NOT** copy the input data files to `/work/`, reference the full path, e.g. `python3 count_gv.py /home/assignments/assignment9/sv.reclassed.filtered.vcf`
- Your submission folder should contain:
  - A completed README.txt
  - Commented scripts:
    - `count_gv.py`
    - `quantify_genotype.py`
    - `violate_MS.py`
  - Figures appropriately scaled with labelled axes and informative titles:
    - `histogram_indels.png`
    - `histogram_deletions.png`
    - `histogram_meis.png`

- Due Wednesday (30th March ‘16) at 10:00 AM


That's all Folks!