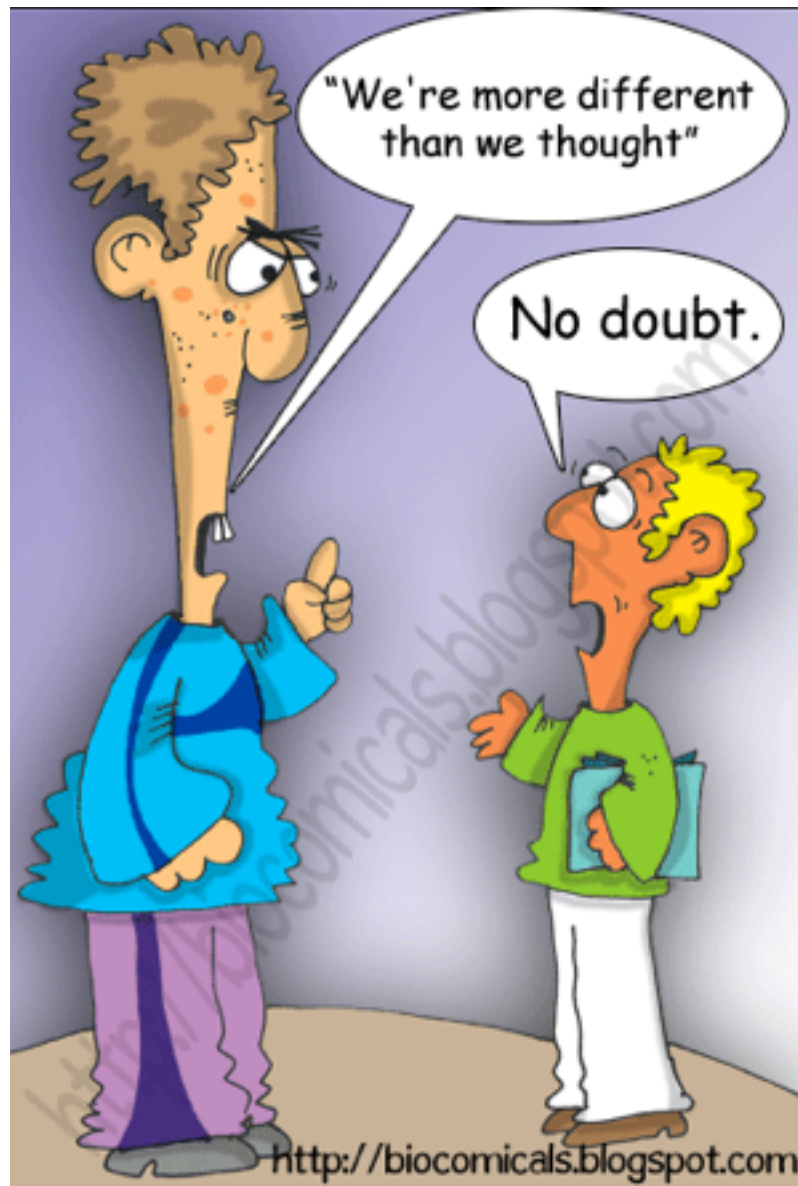


Assignment 9



Mayank Choudhary
Bio5488
25th March 2016

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.reclassified.filtered.vcf`

Output:

Basic VCF parser!! Mendelian violations!!



*

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



First rule first

FIRST RULE OF WORKING WITH VCF FILES



NEVER WRITE YOUR OWN VCF PARSER

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

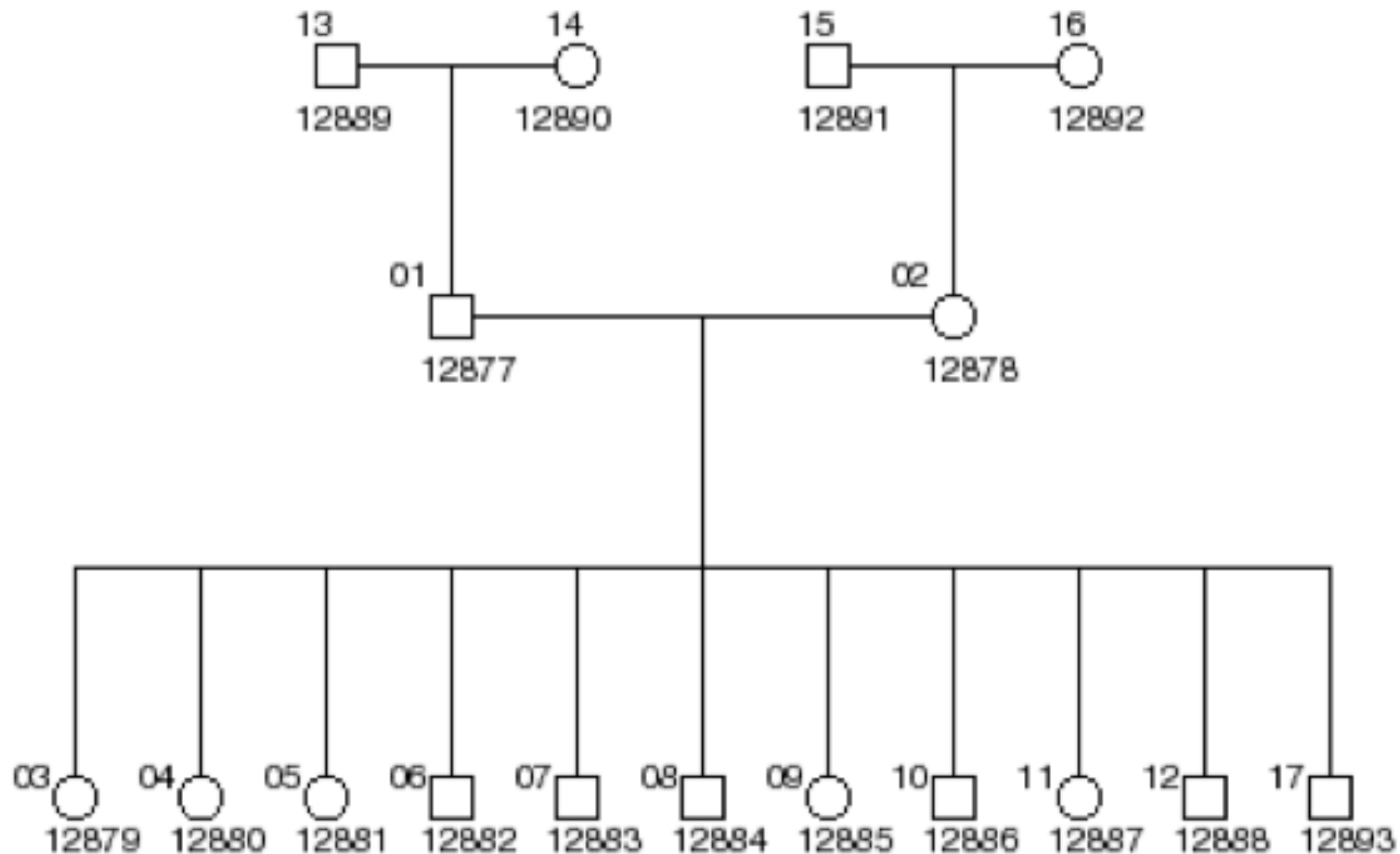


Variant calls from a 17-member family



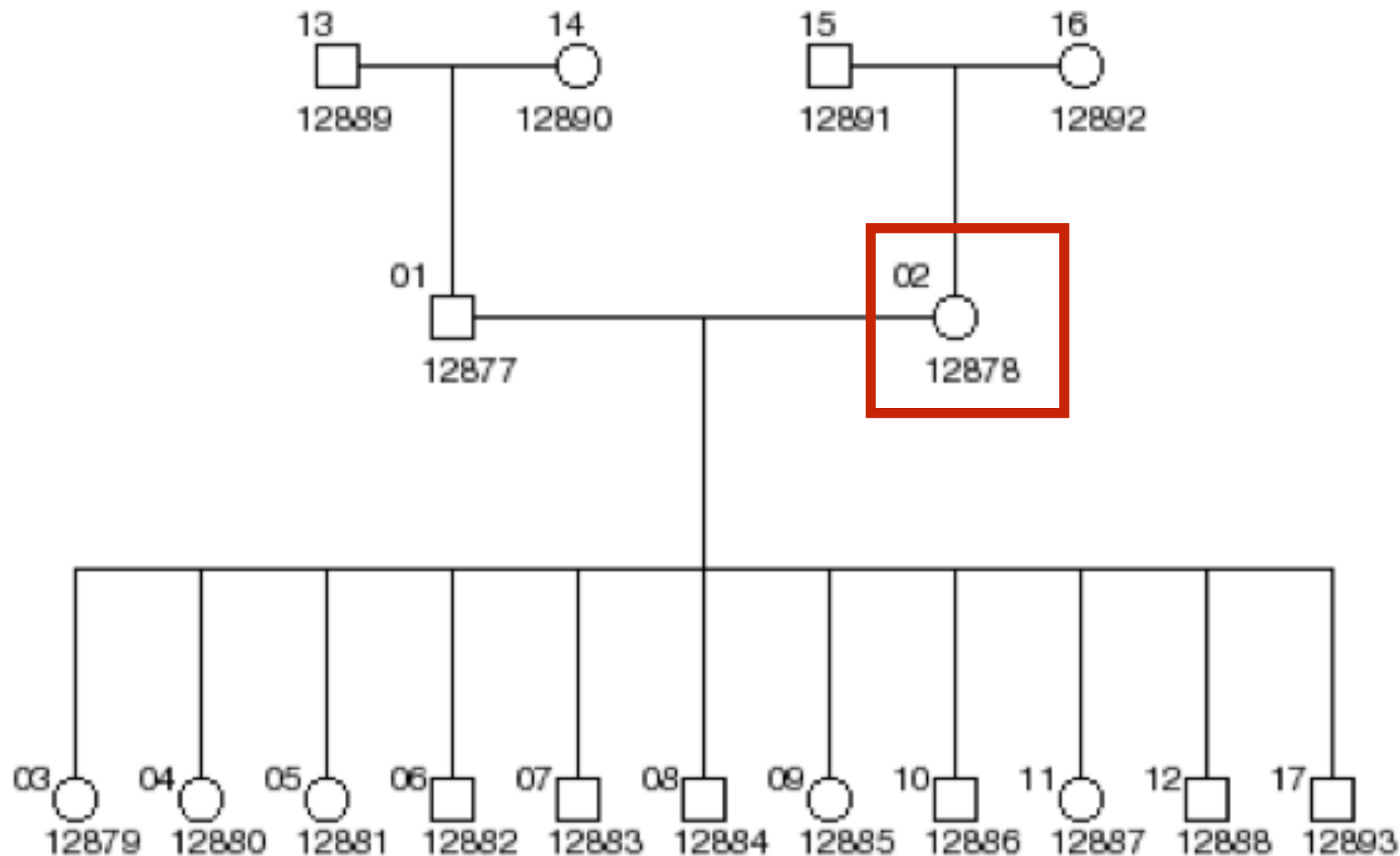
Extremely (re)productive F1s!

CEPH Pedigree 1463



Extremely (re)productive F1s!

CEPH Pedigree 1463

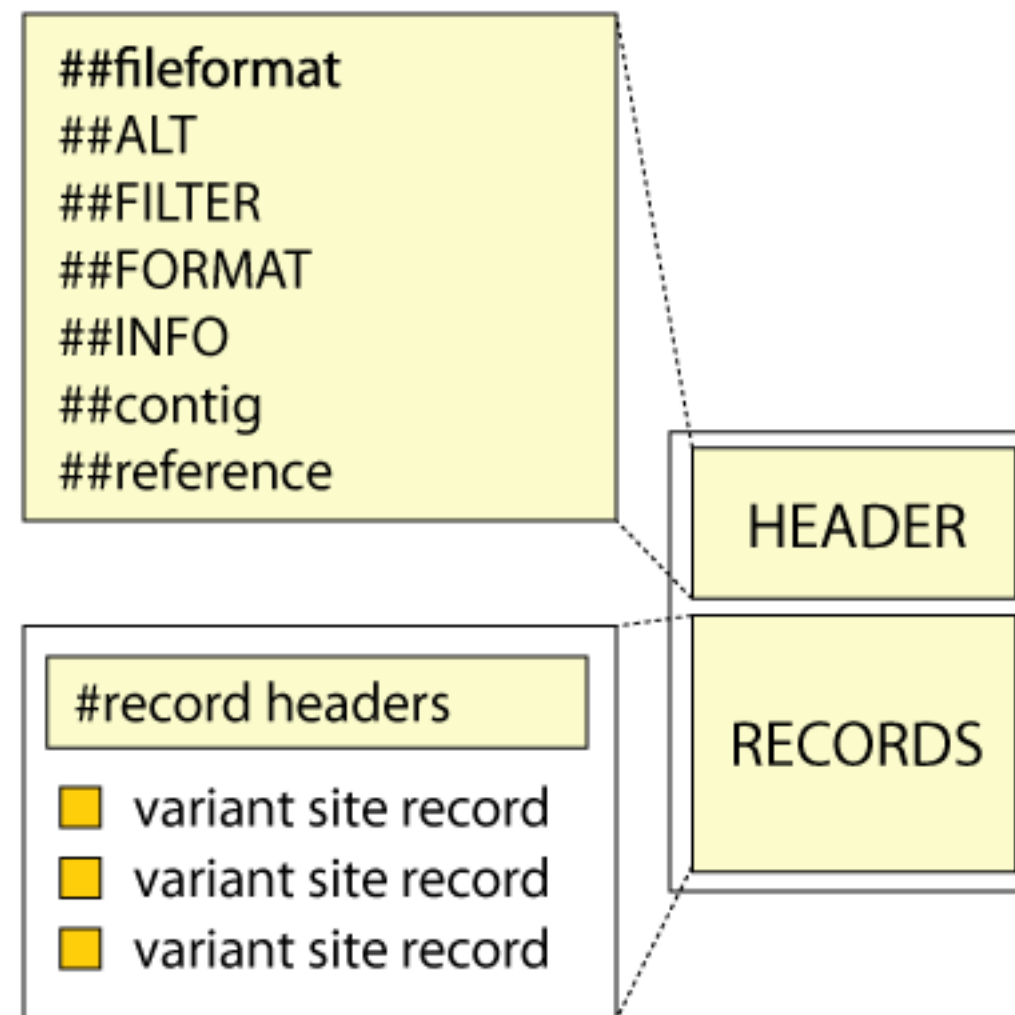


NA12878—the Jonathan Goldsmith of genomes!



Anatomy of a VCF file

Basic structure of a VCF file



Header of a VCF file

```
##FORMAT<ID=A5,Number=A,Type=Integer,Description="Alternate allele split-read observation count, with partial observations recorded fractionally">
##FORMAT<ID=AP,Number=1,Type=Integer,Description="Reference allele paired-end observation count, with partial observations recorded fractionally">
##FORMAT<ID=AR,Number=A,Type=Integer,Description="Alternate allele paired-end observation count, with partial observations recorded fractionally">
##FORMAT<ID=AB,Number=1,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 NA12883 NA12884 NA12885 NA12886 NA12887 NA12888 NA12889 NA12890 NA12891 NA12892 NA12893
[mayank@genomic:/home/assignments/assignment9]$ c

[mayank@genomic:/home/assignments/assignment9]$ head -n51 sv.reclassified.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=1,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=1,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=1,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=1,Type=Integer,Description="Number of pieces of evidence supporting the variant across all samples">
##INFO<ID=PE,Number=1,Type=Integer,Description="Number of paired-end reads supporting the variant across all samples">
##INFO<ID=SR,Number=1,Type=Integer,Description="Number of split reads supporting the variant across all samples">
##INFO<ID=EV,Number=1,Type=String,Description="Type of LUMPY evidence contributing to the variant call">
##INFO<ID=PRPOS,Number=1,Type=String,Description="LUMPY probability curve of the POS breakend">
##INFO<ID=PREND,Number=1,Type=String,Description="LUMPY probability curve of the END breakend">
##INFO<ID=SNAME,Number=1,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=PSQ,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:TAND,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=CQ,Number=1,Type=Float,Description="Genotype quality">
##FORMAT<ID=PQ,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=AQ,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=A5,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=1,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 NA12883 NA12884 NA12885 NA12886 NA12887 NA12888 NA12889 NA12890 NA12891 NA12892 NA12893
[mayank@genomic:/home/assignments/assignment9]$
```


Records in a VCF file

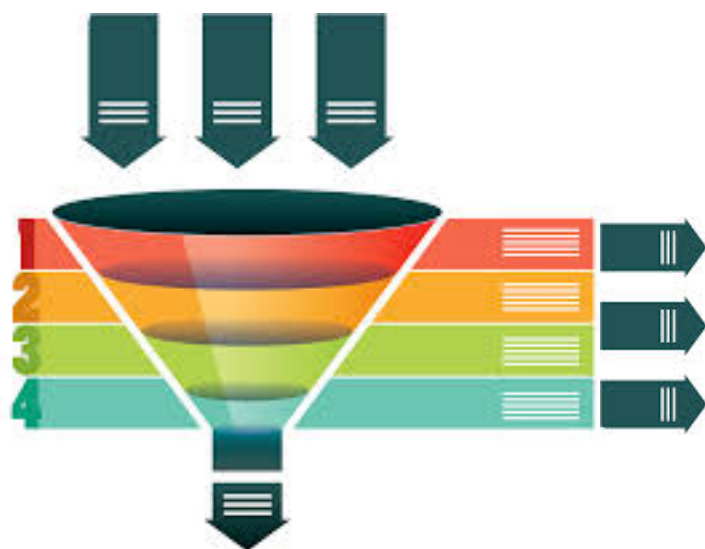
DEL

SNV

INS

DEL

```
##contig=<ID=GL000195.1,length=182896,assembly=b37>
##contig=<ID=GL000212.1,length=186858,assembly=b37>
##contig=<ID=GL000222.1,length=186861,assembly=b37>
##contig=<ID=GL000200.1,length=187035,assembly=b37>
##contig=<ID=GL000193.1,length=189789,assembly=b37>
##contig=<ID=GL000194.1,length=191469,assembly=b37>
##contig=<ID=GL000225.1,length=211173,assembly=b37>
##contig=<ID=GL000192.1,length=547496,assembly=b37>
##reference=file:///gscmnt/gc2719/halllab/genomes/human/GKCh37/1kg_phase1/human_g1k_v37.fasta
##VEP=v76 cache=/gscmnt/gc2719/halllab/src/speedseq/annotations/vep_cache/homo_sapiens/v76_GKCh37_db=
##INFO=<ID=CSQ,Number=,Type=String,Description="Consequence type as predicted by VEP. Format: Consequence[Codons][Amino_acids][Gene][SYMBOL][Feature][EXON][PolyPhen][SIFT][Protein_position][BIOTYPE][LoF][LoF_filter][LoF_flags][LoF_info">
##FORMAT=ID,POS,REF,ALT,QUAL,FILTER,INFO
#CHROM POS ID REF ALT QUAL FILTER INFO
1 10146 . AC A 1105.65 . AC=4;AF=0.118;AN=34;BaseQRankSum=-1.587e+00;ClippingRankSum=0.022;DP=1778;FS=13.924;GQ_MEAN=10.24;GQ_STDDEV=51.73;InbreedingCoeff=-0.2142;MLEAC=11;MLEAF=0.324;MQ=36.57;MQ0=0;MQRankSum=-1.720e-01;NCC=0;QD=7.68;ReadPosRankSum=1.23;SOR=2.268;CSQ=upstream_gene_variant[|]ENSG00000223972|D0X11L1|ENST00000456328|]processed_transcript[|] GT:AD:DP:GQ:PGT:PID:PL 0/0:110:0:110:0:0,2438 0/1:12,61:73:97:482,0,97 0/0:125:0:0,0,2496 1/1:3,33:36:16:192,16,0 0/0:124:0:0,0,2345 0/0:116:0:116:0:0,1531 0/0:90:0:90:0:0,1743 0/0:73:0:73:0:0,0,852 0/0:92:0:92:0:0,1300 0/0:101:0:101:0:0,1984 0/0:107:0:107:0:0,2340
0/0:112:0:112:0:0,2243 0/0:115:0:115:0:0,1956 0/0:64:0:64:0:0,1115 0/1:13,22:35:99:469,0,197 0/0:113:0:113:0:0,1774 0/0:120:0:120:0:0,1959
1 10238 . CCCTAA C 280.31 . AC=1;AF=0.031;AN=32;BaseQRankSum=-1.390e-01;ClippingRankSum=0.694;DP=1726;FS=2.017;GQ_MEAN=10.88;GQ_STDDEV=43.50;InbreedingCoeff=-0.3362;MLEAC=5;MLEAF=0.156;MQ=37.00;MQ0=0;MQRankSum=0.707;NCC=1;QD=0.03;ReadPosRankSum=2.98;SOR=0.730;CSQ=upstream_gene_variant[|]ENSG00000223972|D0X11L1|ENST00000456328|]processed_transcript[|] GT:AD:DP:GQ:PGT:PID:PL 0/0:94:0:94:0:0,1550 0/0:121:0:121:0:0,5750/0:106:0:106:0:0,1101 0/0:92:0:92:0:0,110 0/0:114:0:114 0/0:98:0:98:0:0,454 0/0:74:0:74:0:0,559 0/0:76:0:76:0:0,397 0/0:95:0:95:0:0,525 0/0:83:0:83:0:0,815 0/0:100:0:100:0:0,117:0:117:0:0,1574 0/0:81:0:81:0:0,411 0/1:5,14:19:99:0:1:10238_CCCTAA_C:302,0,174 0/0:133:0:133:0:0,813 0/0:115:0:115:0:0,372
1 10250 . A C 278.63 . AC=1;AF=0.029;AN=34;BaseQRankSum=1.81;ClippingRankSum=-7.870e-01;DP=1712;FS=9.863;GQ_MEAN=15.47;GQ_STDDEV=47.59;InbreedingCoeff=-0.3333;MLEAC=5;MLEAF=0.147;MQ=38.09;MQ0=0;MQRankSum=2.310e-01;NCC=0;QD=14.66;ReadPosRankSum=1.97;SOR=0.613;CSQ=upstream_gene_variant[|]ENSG00000223972|D0X11L1|ENST00000456328|]processed_transcript[|] GT:AD:DP:GQ:PGT:PID:PL 0/0:102:0:102:0:0,985 0/0:117:0:117:0:0,1014 0/0:84:0:84:0:0,850 0/0:118:0:118:0:0,1610 0/0:104:0:104:0:0,1256 0/0:29:0:29:0:0,78,818 0/0:84:0:84:0:0,983 0/1:5,14:19:99:0:1:10238_CCCTAA_C:302,0,105 0/0:134:0:134:0:0,1250/0:116:0:116:0:0,1994
1 10257 . A C 284.48 . AC=1;AF=0.033;AN=30;BaseQRankSum=-6.930e-01;ClippingRankSum=1.07;DP=1764;FS=10.139;GQ_MEAN=12.60;GQ_STDDEV=33.83;InbreedingCoeff=-0.3333;MLEAC=4;MLEAF=0.133;MQ=38.61;MQ0=0;MQRankSum=-1.197e+00;NCC=2;QD=16.73;ReadPosRankSum=0.063;SOR=0.602;CSQ=upstream_gene_variant[|]ENSG00000223972|D0X11L1|ENST00000456328|]processed_transcript[|] GT:AD:DP:GQ:PGT:PID:PL 0/0:107:0:107:0:0,523 0/0:116:0:116:0:0,1024
1097 . /:100:0:100 0/0:94:0:94:0:0,281 0/0:100:0:100:0:0,516 0/0:90:0:90:0:0,632 0/0:93:0:93:0:0,662 0/0:82:0:82:0:0,1021 0/0:100:0:100:0:0,1320 0/0:84:0:84:0:0,554 0/0:123:0:123:0:0,1247 0/0:104:0:104:0:0,1119 0/0:119:0:119:0:0,818 0/1:3,14:17:99:0:1:10238_CCCTAA_C:308,0,111 0/0:143:0:143:0:0,1108 0/0:116:0:116:0:0,1024
1 10291 . C CT 44.42 . AC=1;AF=0.029;AN=34;BaseQRankSum=0.736;ClippingRankSum=0.736;DP=1716;FS=0.000;GQ_MEAN=6.80;GQ_STDDEV=19.85;InbreedingCoeff=-0.3144;MLEAC=4;MLEAF=0.118;MQ=33.71;MQ0=0;MQRankSum=0.736;NCC=0;QD=14.81;ReadPosRankSum=0.736;SOR=1.051;CSQ=upstream_gene_variant[|]ENSG00000223972|D0X11L1|ENST00000456328|]processed_transcript[|] GT:AD:DP:GQ:PGT:PID:PL 0/0:94:0:94:0:0,1238 0/0:103:0:103:0:0,1404 0/0:102:0:102:0:0,1917 0/0:100:0:100:0:0,1035 0/0:93:0:93:0:0,1856 0/0:90:0:90:0:0,1625 0/0:97:0:97:0:0,1270 0/1:1,2:3:31:69,0,31 0/0:93:0:93:0:0,570 0/0:75:0:75:0:0,0,871 0/0:125:0:125:0:0,1393 0/0:90:0:90:0:0,1922 0/0:29:0:29:0:0,78,818 0/0:81:0:81:0:0,1187 0/0:100:0:100:0:0,2066 0/0:127:0:127:0:0,1664 0/0:115:0:115:0:0,1523
1 10354 . C A 255.13 . AC=3;AF=0.375;AN=8;BaseQRankSum=0.687;ClippingRankSum=-1.399e+00;DP=2509;FS=2.701;GQ_MEAN=91.25;GQ_STDDEV=65.78;MLEAC=3;MLEAF=0.375;MQ=30.24;MQ0=0;MQRankSum=-3.650e-01;NCC=13;QD=6.71;ReadPosRankSum=0.731;SOR=1.259;CSQ=upstream_gene_variant[|]ENSG00000223972|D0X11L1|ENST00000456328|]processed_transcript[|] GT:AD:DP:GQ:PGT:PID:PL 0/1:20,0:28:99:160,0,438 ./:109,0:109 ./:100,0:100 0/1:6,4:10:99:109,0,132 ./:157,0:157 ./:173,0:173 ./:195,0:195 ./:141,0:141 0/0:29:0:29:0:0,78,818 ./:144,0:144 ./:165,0:165 ./:159,0:159 ./:216,0:216
1 10403 . ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC A 4109.05 . AC=9;AF=0.265;AN=34;BaseQRankSum=2.19;ClippingRankSum=-4.150e-01;DP=2253;FS=8.222;GQ_MEAN=274.24;GQ_STDDEV=289.81;InbreedingCoeff=-0.4783;MLEAC=11;MLEAF=0.324;MQ=36.71;MQ0=0;MQRankSum=0.677;NCC=0;QD=0.93;ReadPosRankSum=-2.44e+00;SOR=1.331;CSQ=upstream_gene_variant[|]ENSG00000223972|D0X11L1|ENST00000456328|]processed_transcript[|] GT:AD:DP:GQ:PGT:PID:PL 0/1:25,13:38:99:1393,0,1070 0/0:209:0:209:99:111,1000 0/1:24,5:29:99:129,0,991 0/1:20,20:40:99:681,0,801 0/0:202:0:202:99:120,1000 0/0:264:0:264:0:0,5559 0/0:144:0:144:0:0,2037 0/1:0,5:13:99:0:1:10403_ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC_A:167,0,271 0/1:37,9:46:99:266,0,1157 0/1:19,17:36:99:0:1:10403_ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC_A:537,0,1442 0/1:29,16:45:99:515,0,806 0/0:187:0:187:0:0,3442 0/0:29:0:29:0:0,78,818 0/0:40:0:40:99:101,1222 0/0:247:0:247:0:0,4041 0/1:26,26:52:99:0:1:10403_ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC_A:968,0,3783 0/1:41,18:59:99:596,0,1733
1 10409 . ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC A 711.05 . AC=4;AF=0.118;AN=34;BaseQRankSum=1.57;ClippingRankSum=0.442;DP=1912;FS=5.151;GQ_MEAN=112.35;GQ_STDDEV=117.13;InbreedingCoeff=-0.1323;MLEAC=4;MLEAF=0.118;MQ=36.13;MQ0=0;MQRankSum=0.324;NCC=0;QD=0.46;ReadPosRankSum=-3.230e-01;SOR=1.118;CSQ=upstream_gene_variant[|]ENSG00000223972|D0X11L1|ENST00000456328|]processed_transcript[|] GT:AD:DP:GQ:PGT:PID:PL 0/1:32,3:35:15:15,0,1290/0:89:0:209:99:110,111,1000 0/0:24:0:29:75:10,75,1085 0/0:28:0:48:75:10,75,939 0/0:202:0:202:99:120,1000 0/0:237:0:237:91:10,91,1000 0/0:147:0:147:99:10,102,1530 0/0:8:0:13:26:0:1:10403_ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC_A:0,26,313 0/1:36,4:40:61:61,0,1140 0/0:30:1:31:46:10:10403_ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC_A:0,46,1660 0/1:36,8:44:99:154,0,1257 0/1:17,16:33:99:0:1:10409_ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC_A:544,0,1004 0/0:41:1,42:60:60,722 0/0:40:0:40:99:10,103,1226 0/0:231:0:231:99:120,1000 0/0:26:0:52:81:0:1:10403_ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC_A:0,81,3953 0/0:41:0:59:99:10,126,1915
```

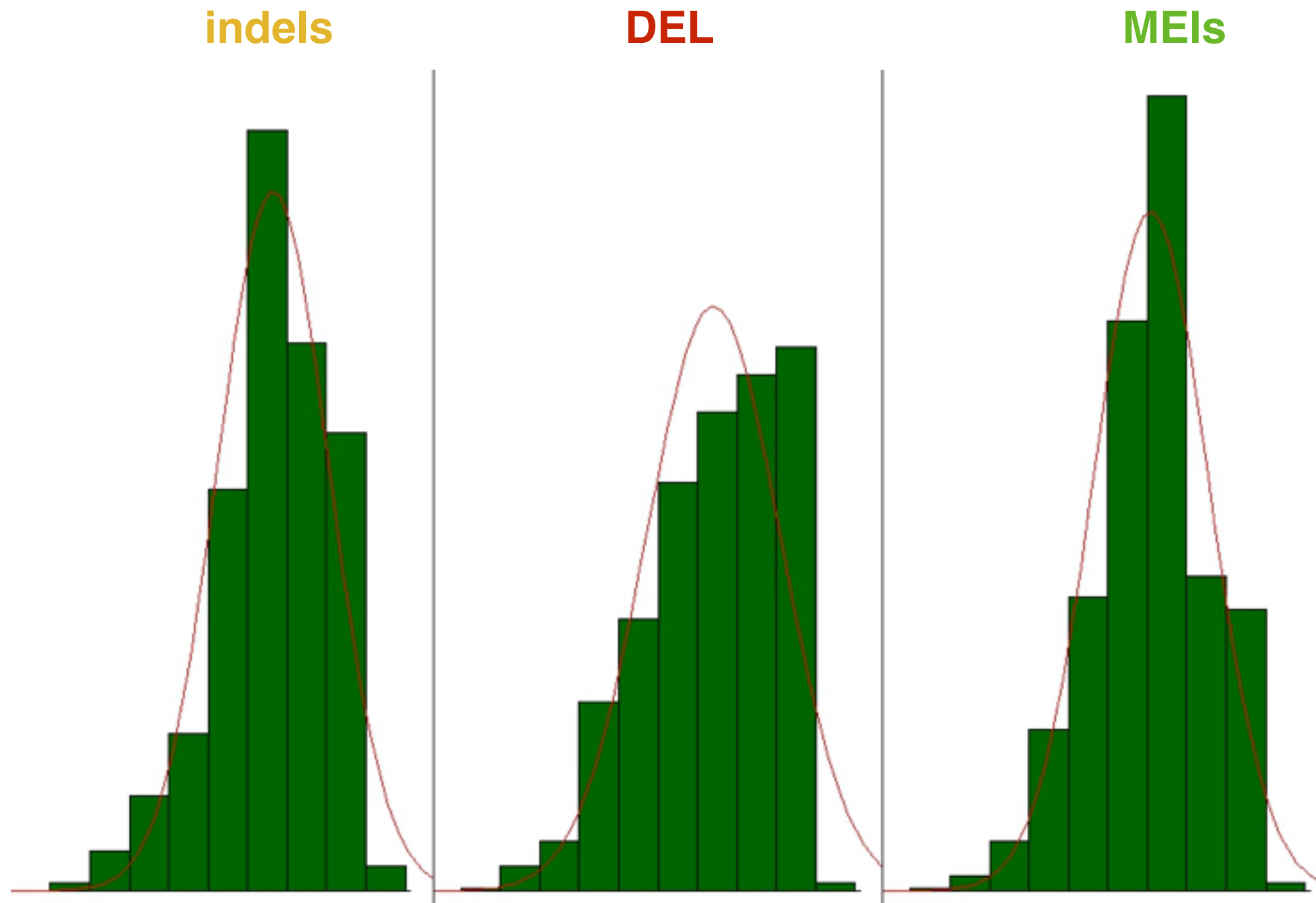


Profiling counts

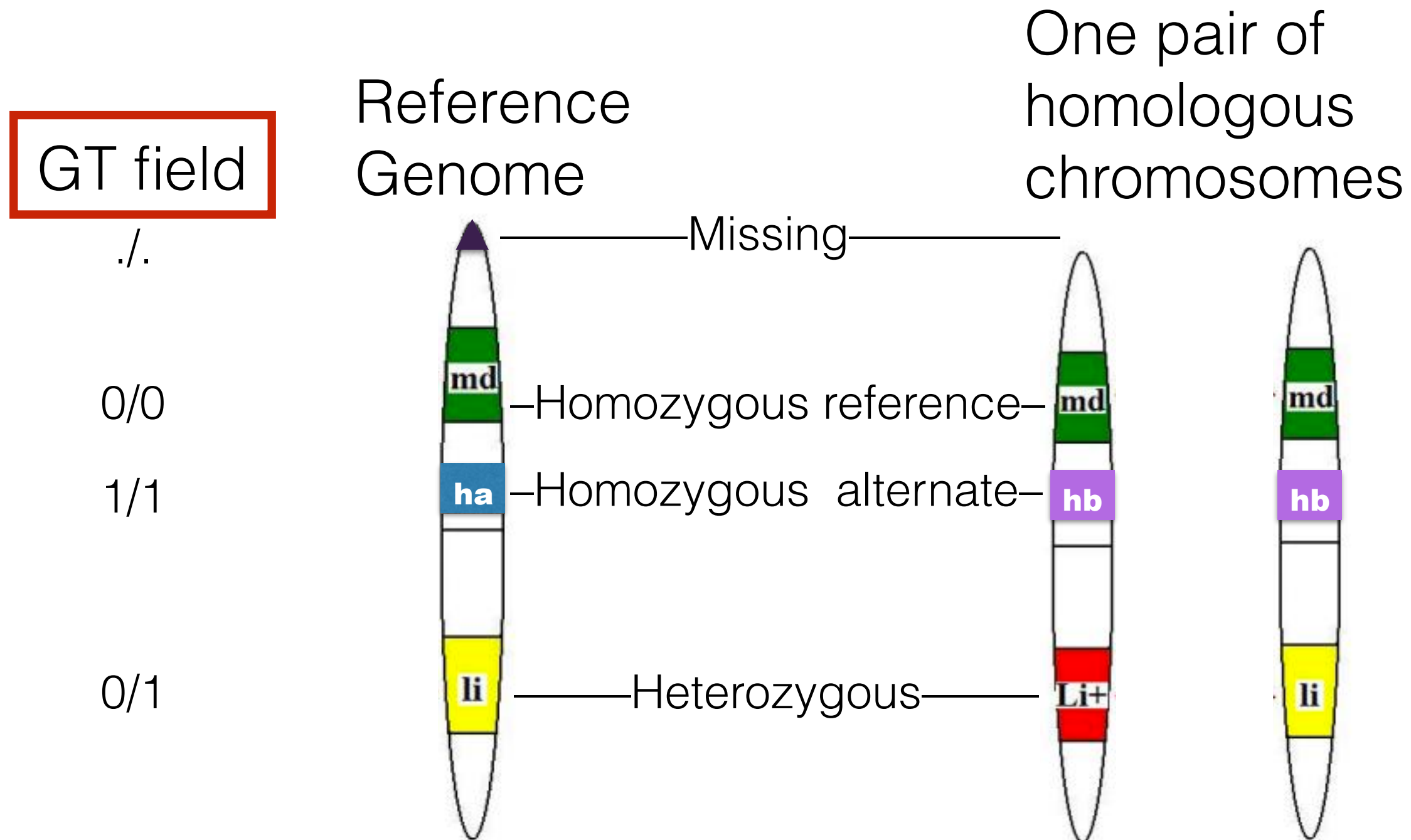
Class of genome variation	count
SNVs
indels
DEL
DUP	...
INV	..
MEIs	...
BNDs

Total GV

Plot the size distributions



Zygosity explained

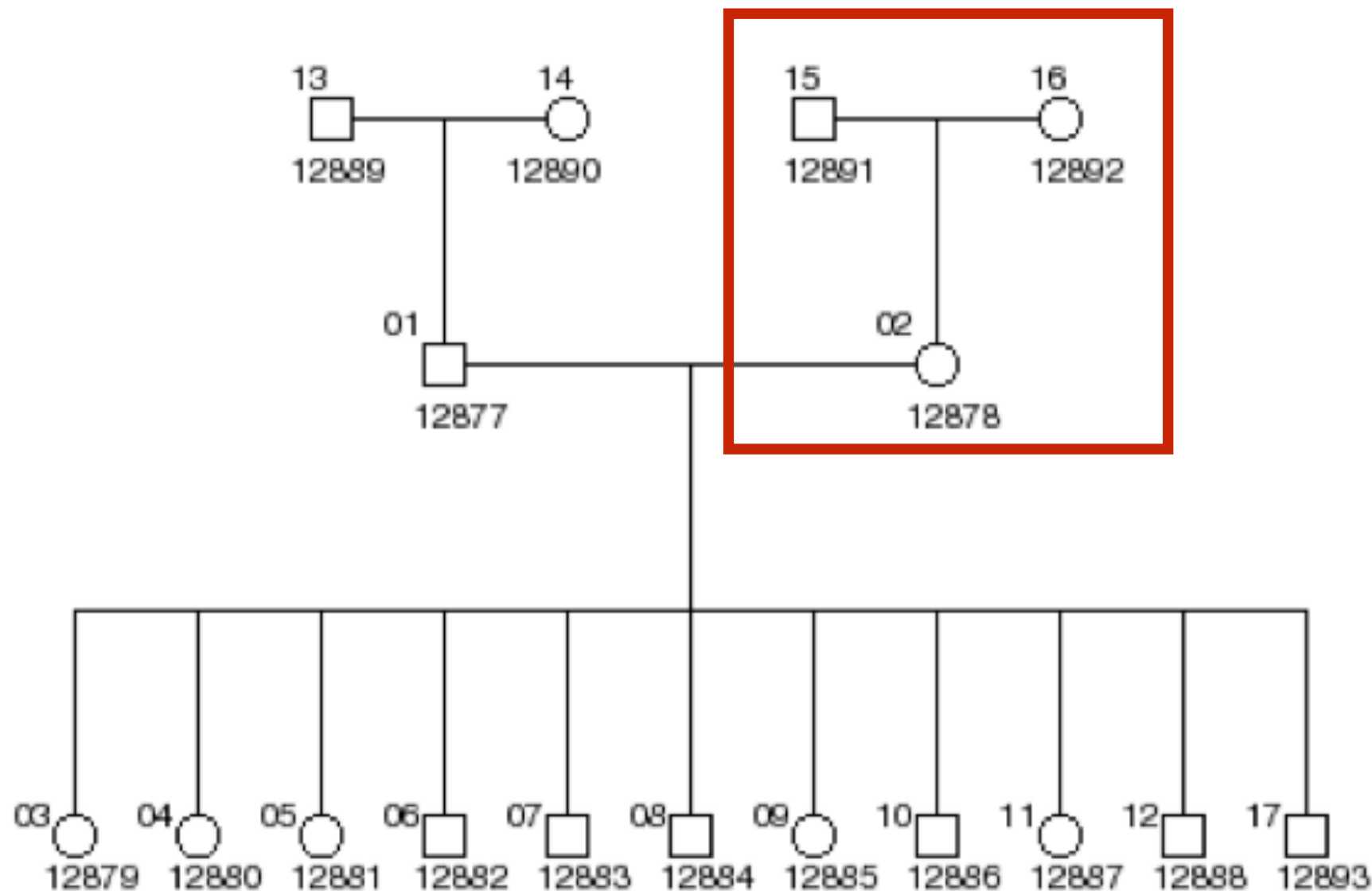


A wise (now dead) man once said...

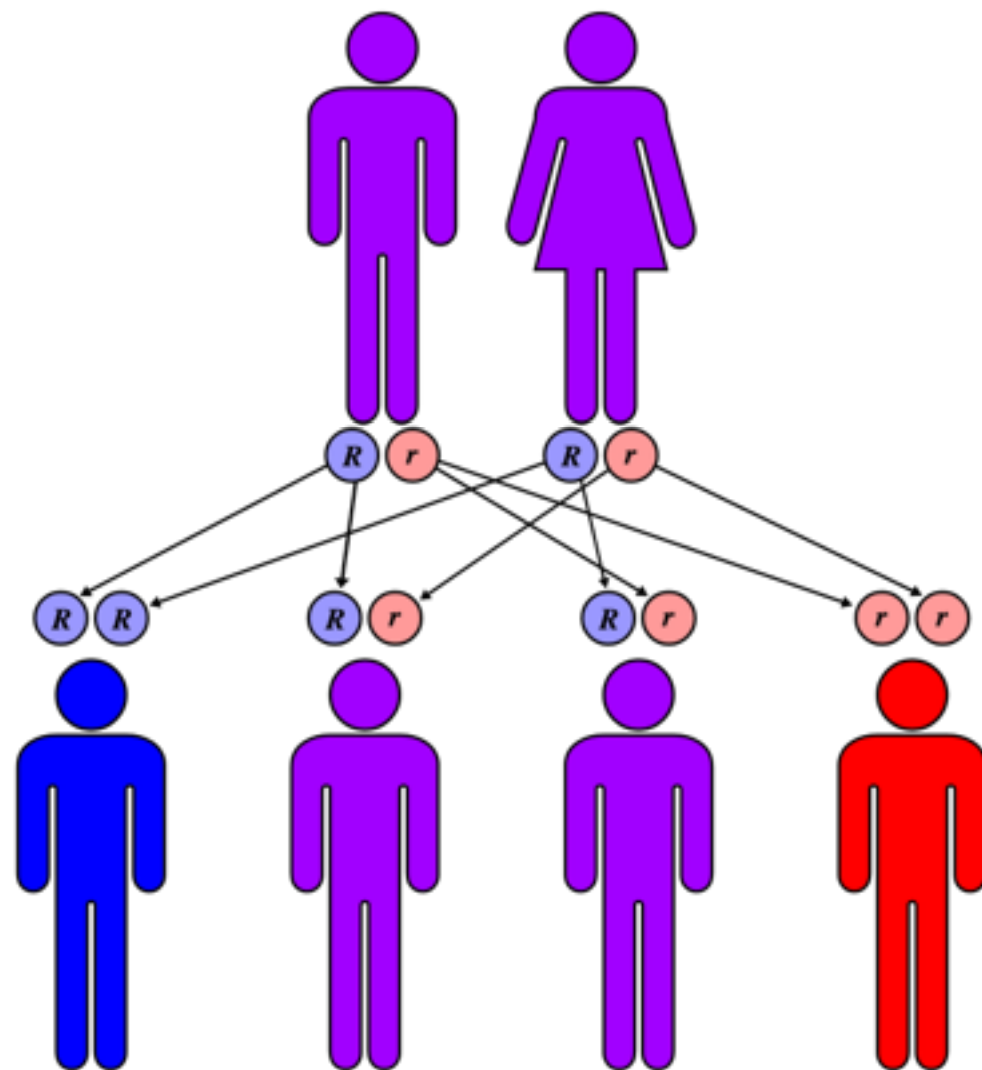


Trio analysis to look for violations of Mendel's Law of Segregation

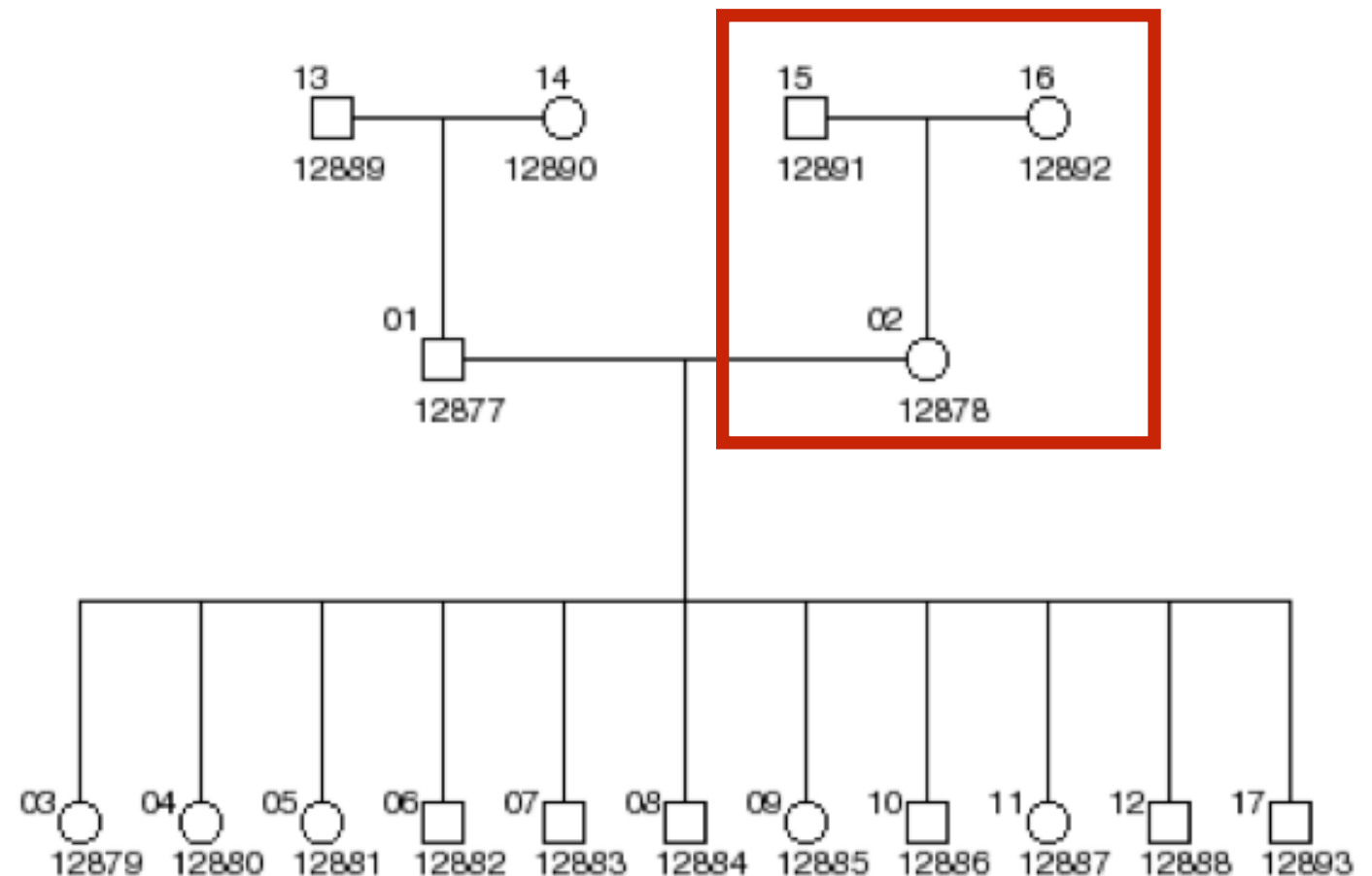
CEPH Pedigree 1463



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



CEPH Pedigree 1463



WHY?



BUT WHERE'S QUESTION 9?



GQ—Genotype Quality



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

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

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

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.

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.reclassified.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

