Exome Sequencing and Disease Gene Search

Erzurumluoglu AM, Rodriguez S, Shihab HA, Baird D, Richardson TG, Day IN, Gaunt TR. **Identifying Highly Penetrant Disease Causal Mutations Using Next Generation Sequencing: Guide to Whole Process**. Biomed Res Int. 2015;2015:923491. doi: 10.1155/2015/923491. Epub 2015 Apr 6. Review. PubMed PMID: 26106619; PubMed Central PMCID: PMC4461748.

Recent technological advances have created challenges for geneticists and a need to adapt to a wide range of new bioinformatics tools and an expanding wealth of publicly available data (e.g., mutation databases, and software). **This wide range of methods and a diversity of file formats used in sequence analysis is a significant issue**, with a considerable amount of time spent before anyone can even attempt to analyse the genetic basis of human disorders. Another point to consider that is although many possess "just enough" knowledge to analyse their data, **they do not make full use of the tools and databases that are available and also do not fully understand how their data was created**. The primary aim of this review is to document some of the key approaches and provide an analysis schema to make the analysis process more efficient and reliable **in the context of discovering highly penetrant causal mutations/genes**. This review will also compare the methods used to identify highly penetrant variants when data is obtained **from consanguineous individuals** as opposed to **nonconsanguineous**; and when **Mendelian disorders** are analysed as opposed to **common-complex disorders**.

Variant Call Format (VCF)

- Specifies the text file used in bioinformatics for storing gene sequence variations.
- The format has been developed with the advent of large-scale genotyping and DNA sequencing projects, such as the 1000 Genomes Project.
- By using the variant call format only the variations need to be stored along with a reference genome.

```
##fileformat=VCFv4.0 ##fileDate=20090805 ##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36 ##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10"> ##FILTER=<ID=s50,Description="Less
than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

INFO, FILTER, FORMAT:

File meta-information is included after the ## string, often as key=value pairs.

Genotype fields specified in the FORMAT field should be described as follows:

##FORMAT=<ID=ID,Number=number,Type=type,Description=”description”>

See

VCF (Variant Call Format) version 4.0

http://www.1000genomes.org/wiki/Analysis/vcf4.0

```
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G     A      29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T     A      3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A     G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T     .      47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT  G,GTACT 50   PASS   NS=3;DP=9;AA=G                   GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

The header line names the 8 fixed, mandatory columns. These columns are as follows: #CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO

The last FORMAT column is optional.

The allele values are 0 for the reference allele (what is in the reference sequence), 1 for the first allele listed in ALT, 2 for the second allele list in ALT and so on.

- / : genotype unphased
- | : genotype phased

*Phased means I know not only the genotypes but which chromosome each genotype call came from. This lets you interpret which sets of genotypes are being inherited together

INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data]. Arbitrary keys are permitted, although the following sub-fields are reserved (albeit optional):

- AA ancestral allele
- AC allele count in genotypes, for each ALT allele, in the same order as listed
- AF allele frequency for each ALT allele in the same order as listed: use this when estimated from primary data, not called genotypes
- AN total number of alleles in called genotypes
- BQ RMS base quality at this position
- CIGAR cigar string describing how to align an alternate allele to the reference allele
- DB dbSNP membership
- DP combined depth across samples, e.g. DP=154
- END end position of the variant described in this record (esp. for CNVs)
- H2 membership in hapmap2
- MQ RMS mapping quality, e.g. MQ=52
- MQ0 Number of MAPQ == 0 reads covering this record
- NS Number of samples with data
- SB strand bias at this position
- SOMATIC indicates that the record is a somatic mutation, for cancer

If genotype information is present, then the same types of data must be present for all samples. First a FORMAT field is given specifying the data types and order. This is followed by one field per sample, with the colon-separated data in this field corresponding to the types specified in the format. The first sub-field must always be the genotype (GT).

As with the INFO field, there are several common, reserved keywords that are standards across the community:

•GT genotype, encoded as alleles values separated by either of ”/” or “|”, …. All samples must have GT call information; if a call cannot be made for a sample at a given locus, ”.” must be specified for each missing allele in the GT field (for example ./. for a diploid). The meanings of the separators are:

- / : genotype unphased
- | : genotype phased

•DP read depth at this position for this sample (Integer)

•FT sample genotype filter indicating if this genotype was “called” (similar in concept to the FILTER field). …

•GL : three floating point log10-scaled likelihoods for AA,AB,BB genotypes where A=ref and B=alt; not applicable if site is not biallelic. For example: GT:GL 0/1:-323.03,-99.29,-802.53 (Numeric)

•GQ genotype quality, encoded as a phred quality -10log_10p(genotype call is wrong) (Numeric)

•HQ haplotype qualities, two phred qualities comma separated

This example shows in order

- a good simple SNP,
- a possible SNP that has been filtered out because its quality is below 10,
- a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error),
- a site that is called monomorphic reference (i.e. with no alternate alleles),
- and a microsatellite with two alternative alleles, one a deletion of 3 bases (TCT), and the other an insertion of one base (A).

Genotype data are given for three samples, two of which are phased and the third unphased, with per sample genotype quality, depth and haplotype qualities (the latter only for the phased samples) given as well as the genotypes. The microsatellite calls are unphased.