Tools and Algorithms in Bioinformatics

GCBA815, Fall 2013

Week-13: NextGen Sequence Analysis

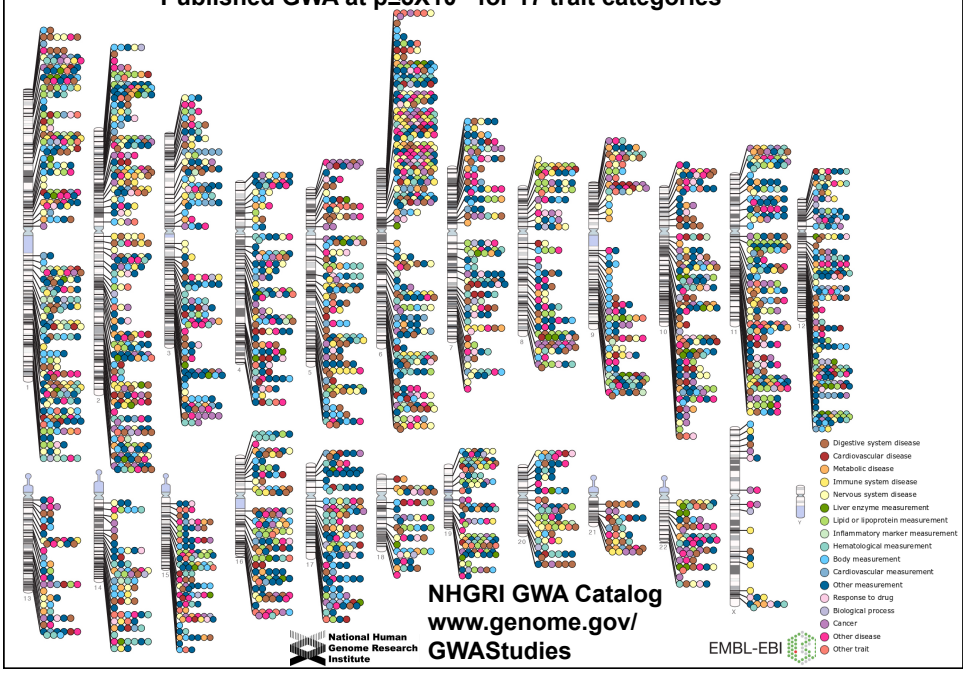
Demonstrators: Suleyman Vural, Li You, Sanjit Pandey

Babu Guda

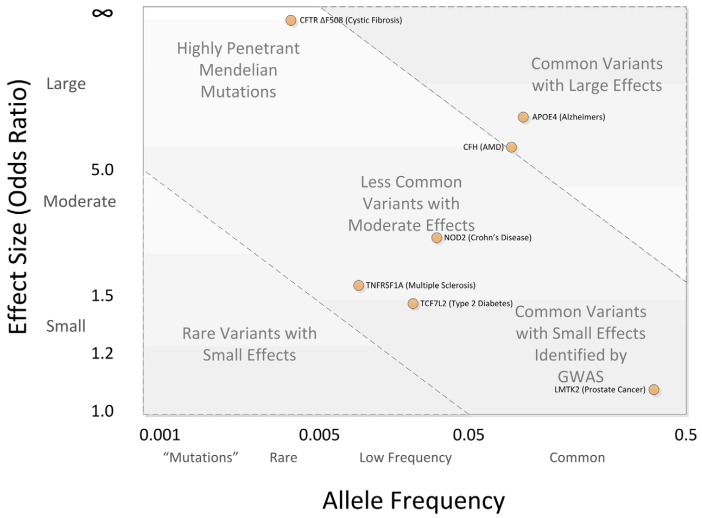
Department of Genetics, Cell Biology and Anatomy

University of Nebraska Medical Center

Published Genome-Wide Associations through 12/2012
Published GWA at p≤5X10^{-8} for 17 trait categories



Allele Frequency vs Effect size of rare and common variants

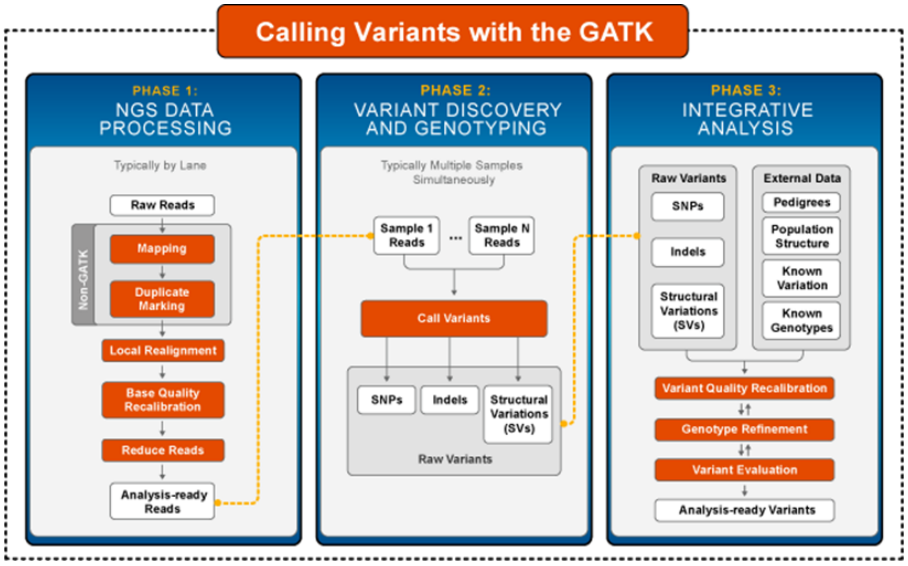


IUPAC codes for nucleotides

Code	Definition	Meaning
A	Adenine	**A**
C	Cytosine	**C**
G	Guanine	**G**
T	Thymine	**T**
R	AG	pu**R**ine
Y	CT	p**Y**rimidine
K	GT	**K**eto
M	AC	a**M**ino
S	GC	**S**trong
W	AT	**W**eak
B	CGT	Not A
D	AGT	Not C
H	ACT	Not G
V	ACG	Not T
N	AGCT	a**N**y

NGS Applications in Biosciences

- Genome
  - Exome sequencing
    - Clinical sequencing, personalized medicine
  - Targeted genome sequencing (Ex: Ion torrent amplicons)
  - Whole genome sequencing
- Transcriptome
  - Whole transcriptome analysis
  - Small RNA analysis (siRNA, lncRNA, miRNA)
  - Gene expression profiling for selected target genes
- Metagenome
  - Sequencing together the genomes of a mixture of species
  - Example: Human gut microbiota or environmental samples
- Epigenome
  - Chromatin Immunoprecipitation Sequencing (ChIP-Seq)
  - Methylation and chromatin remodeling studies



Different file types in NGS analysis

- Fastq file – generated by the sequencer, contains NGS reads
- SAM file – Sequence Alignment/Map (generated by aligning the NGS reads with the reference genome)
- BAM file – Binary version of the SAM file (SAMtools are used to manipulate SAM/BAM files)
- GFF file – General Feature Format used to hold genome annotation (chromosome, strand, frame, exon, CDS, etc.)
- GTF file – Gene Transfer Format (Also contains all the info as in GFF and in addition contains gene annotation information)
- VCF file – Variant Call Format (used to store variant data such as SNPs, InDels, short structural rearrangements)

FASTQ format:

FASTQ is based on the popular FASTA format for sequences

FASTA format
>sequence_ID; header in one line
AGTTGTAGTCCGTGATAGTCGGATCGG

FASTQ format provides additional information that includes the quality score
@20FUKAAXX100202:1:64:10634:114560/1
TTGTATTTTTAGTAGAGACGGAGTTTCGCCATGTTGGTCAGGCTGGCCTCGAATTCCTGACCTCAAGTGATCCGCCCGCCTCGGCCTCCCAACGTTTTGG
+
?=@7=>B==:;BB?<B?=8539<6?6>8>=BB<<B=08:9@5;:A@@?@9:BAAA<?;8;@AC@BBBBBA?<9-@B@;CAA77<:BEB<BB@07?@=<?84

ASCII code for Quality score (Phred score, ranges from 0-50)

ASCII code for Quality score (in the increasing order; ! is the worst and ~ is the best

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```


Interpretation of Quality Score (Phred score)

Phred score (Q) vs Error probability (P)

$$Q = -10\log_{10} P$$

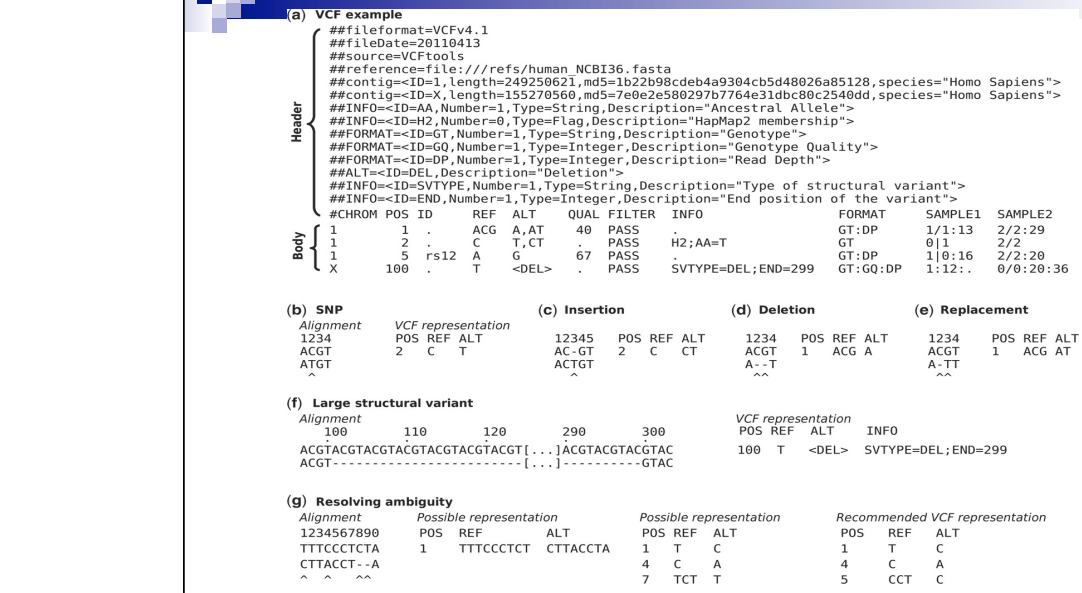
Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

SAM File Format

```
@HD    VN:1.0
@SQ    SN:chr20 LN:62435964
@RG    ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG    ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99  chr20 28833 20 10M1D25M = 28993 195 \
       AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
       NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
       ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<<;<<<<7;:<<<6;<<<<<<<<<<<7<<<< \
       MF:i:18 RG:Z:L2
```

@HD = Header
@SQ = Sequence Dictionary
    LN=length of sequence
@RG= Read Group
    ID=unique read group identifier'
    PU=Platform Unit
    LB=Library
    SM=Sample



Danecek P et al. Bioinformatics 2011;27:2156-2158