# Identify the most common germline ERBB-family SNPs in HER-2 positive Breast Cancer patients via high depth NGS.

By :

Manar Hashem Taha

**Abstract** :

Breast cancer (BC) is the most common malignancy in women with over 25% of all cancers being diagnosed as BC in 2018. HER2-positive BC accounts for around 20% of all human BCs and HER2 overexpression is associated with poor prognosis and an aggressive phenotype. Trastuzumab is a monoclonal antibody targeted to HER2, has well established efficacy in the treatment of HER2-positive BC. However, a significant proportion of patients with the disease have tumors that initially do not respond or that acquire resistance to trastuzumab after an initial period of response . ERBB-family genes which encode the HER family of proteins EGFR, HER2, HER3 and HER4 are commonly studied in HER2-positive BC, and some studies have identified the role of HER2 SNPs in response to trastuzumab .Here we want to identify the most common germline ERBB-family SNPs in HER-2 positive BC patients by via high depth NGS.

## List of Figures

Contents :

- Chapter 2: Download data (Samples / reference genome / VCF annotation file)
- Chapter 3: Methods ( Sequencing analysis ) and Result
- Chapter 4 : Future Work
- References

# Introduction

Breast cancer (BC) remains the most common form of malignancy in women with over 25% of all cancers being diagnosed as BC in 2012 [1]. In HER2-positive BC, which accounts for approximately 20% of all human BCs, HER2 gene amplification and overexpression is associated with an aggressive phenotype and poor prognosis [2]. Trastuzumab, a monoclonal antibody targeted to HER2, has well established efficacy in the treatment of HER2- positive BC [3, 4]. However, a significant proportion of patients with the disease have tumors that initially do not respond or that acquire resistance to trastuzumab after an initial period of response [3, 4]. Many potential mechanisms of trastuzumab resistance in HER2-positive BC have been proposed which have been discussed in detail by us and others [2, 5]; including altered intracellular signaling involving loss of PTEN, reduced p27kip1, increased PI3K/Akt activity (e.g. PIK3CA mutations) or altered signaling via non-HER family receptor tyrosine kinases such as IGF1R . However, few studies have been conducted to understand the role of innate resistance to trastuzumab. The advent of next generation sequencing (NGS) has allowed researchers access to the information stored in the genome. Whilst much focus has been targeted towards the somatic mutations in cancer, less attention has been focused on the role of germline single nucleotide polymorphisms (SNPs) and their role in cancer development and therapy response. In fact, recent studies have identified that SNPs can be biomarkers of the likelihood of developing cancer and several have been implicated in targeted therapy response and resistance [6]. ERBB-family genes which encode the HER family of proteins EGFR, HER2, HER3 and HER4 are commonly studied in HER2-positive BC, and some studies have identified the role of HER2 SNPs in response to trastuzumab [7]. Here we have determined the frequency of germline ERBB-family SNPs in HER-2 positive BC patients by NGS and correlated their genotype with the progression of HER2-positive BC and trastuzumab response.

# DownloadData

**1- Download dataset:**

Download three samples whole exome sequencing of Irish HER2+ breast cancer patients from SRA SRR7309332, SRR7309338, SRR7309325 available at (https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=475755) The library comprised the ERBB-family genes and included 132 regions of coding exons of the ERBB family genes, and the layout is paired.

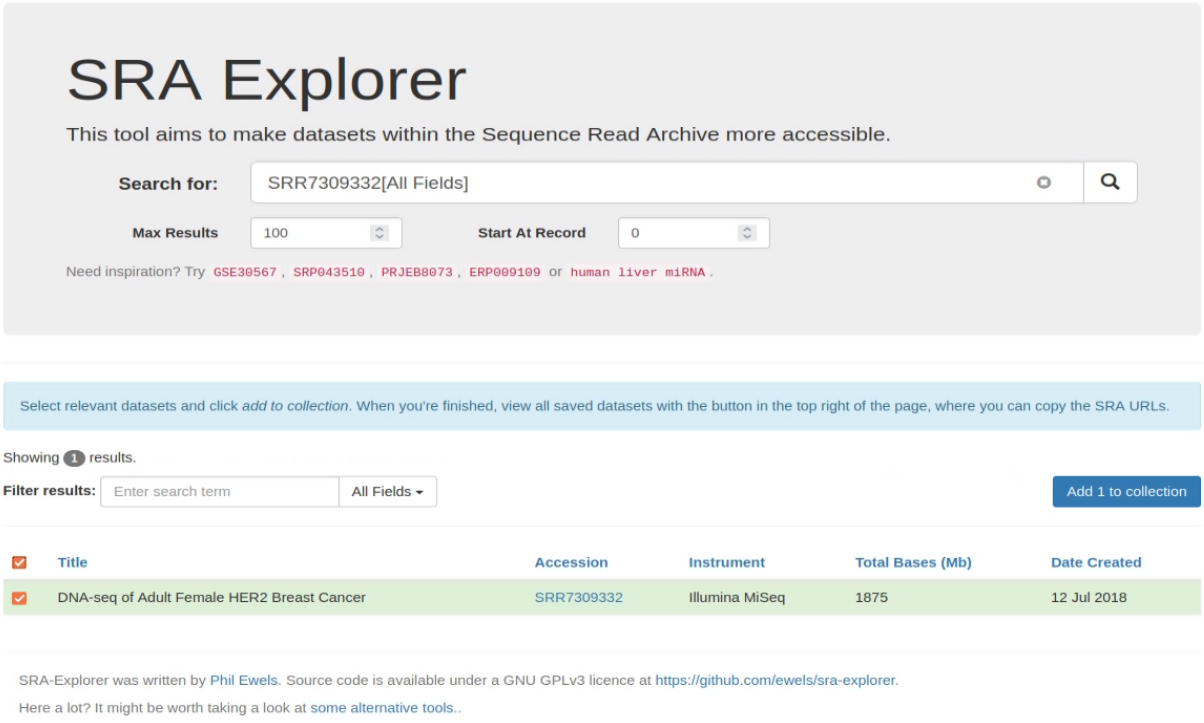


Figure1:SRA Explorer

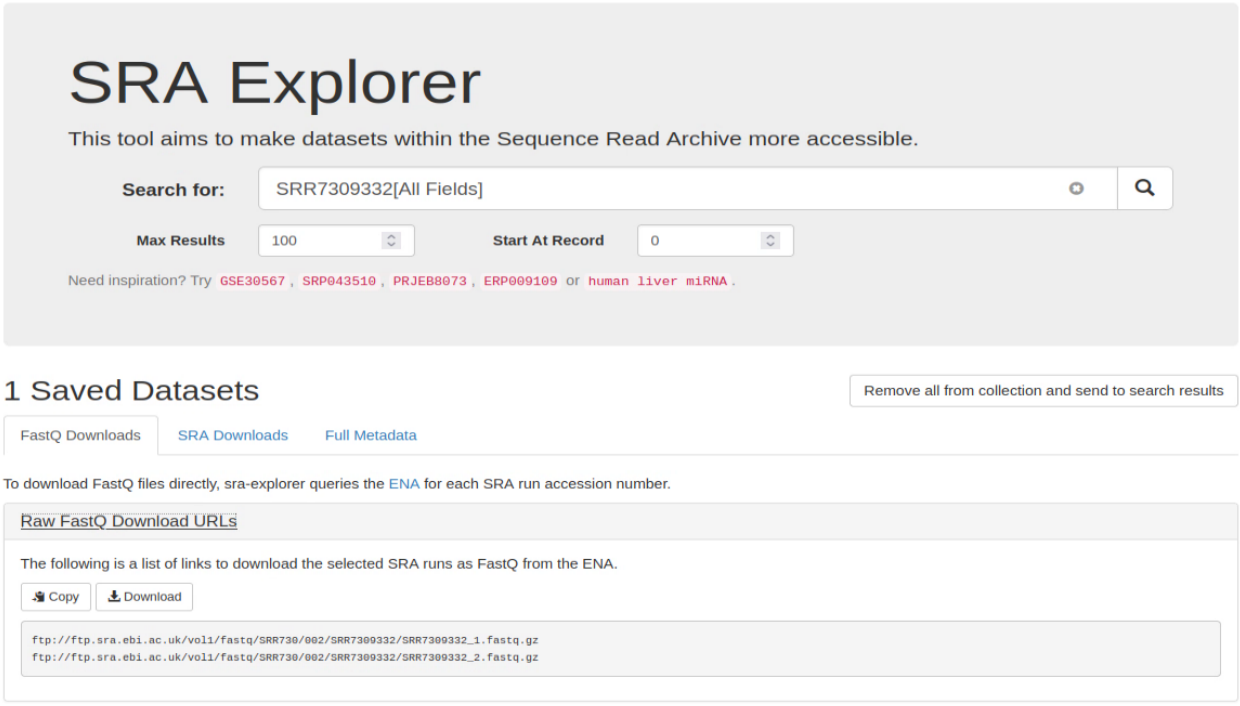


Figure2: Downloads links from SRA Explorer

**2- Download reference genome:**

**-** At first, we downloaded the whole genome however my lab crashed, therefore we thought about selecting the chromosomes that represent ERBB family genes (2,7,12,17) to be a reference and concatenate them in a single Fasta file. Available at : (ftp://ftp.ensembl.org/pub/release-99/fasta/homo_sapiens/dna/)

Second issue, there are multiple fasta file formats to download; We were confused between which format is the best for alignment. T**he files with "sm" in the name are "soft masked" or without masking is the best and any file with "rm" in the name should be avoided.**

The answer available in these websites:

(http://genomespot.blogspot.com/2015/06/mapping-ngs-data-which-genome-version.html)

**3- Download VCF annotation file:**

**We downloaded the VCF files for chromosomes 2, 7, 12, 17 and concatenate them in one file.** The data available in (ftp://ftp.ensembl.org/pub/release-99/variation/vcf/homo_sapiens/)

As in fasta there are multiple VCF format and we chose (homo_sapiens-chr*.vcf.gz) based on a readme file which recommends it for All germline variations from the current Ensembl release.

## Methods ( Sequencing analysis )

# Create conda environment .Install bioconda if you don't have it already.Install the softwares .

### 1- Check Quality of the data and Trimming in case of bad quality

The quality of data was checked using fastqc and resulted in that the data quality were very good so we didn't do trimming .

General Statistics

Copy table | Configure Columns | Plot | Showing 6/6 rows and 3/5 columns.

| Sample Name | % Dups | % GC | M Seqs |
|---|---|---|---|
| SRR7309325_DNA-seq_of_Adult_Female_HER2_Breast_Cancer_1 | 46.9% | 45% | 1.2 |
| SRR7309325_DNA-seq_of_Adult_Female_HER2_Breast_Cancer_2 | 45.2% | 45% | 1.2 |
| SRR7309332_DNA-seq_of_Adult_Female_HER2_Breast_Cancer_1 | 45.8% | 46% | 1.2 |
| SRR7309332_DNA-seq_of_Adult_Female_HER2_Breast_Cancer_2 | 45.9% | 45% | 1.2 |
| SRR7309338_DNA-seq_of_Adult_Female_HER2_Breast_Cancer_1 | 4.9% | 41% | 1.2 |
| SRR7309338_DNA-seq_of_Adult_Female_HER2_Breast_Cancer_2 | 4.6% | 41% | 1.2 |

Figure 3 : QC  General Statistics

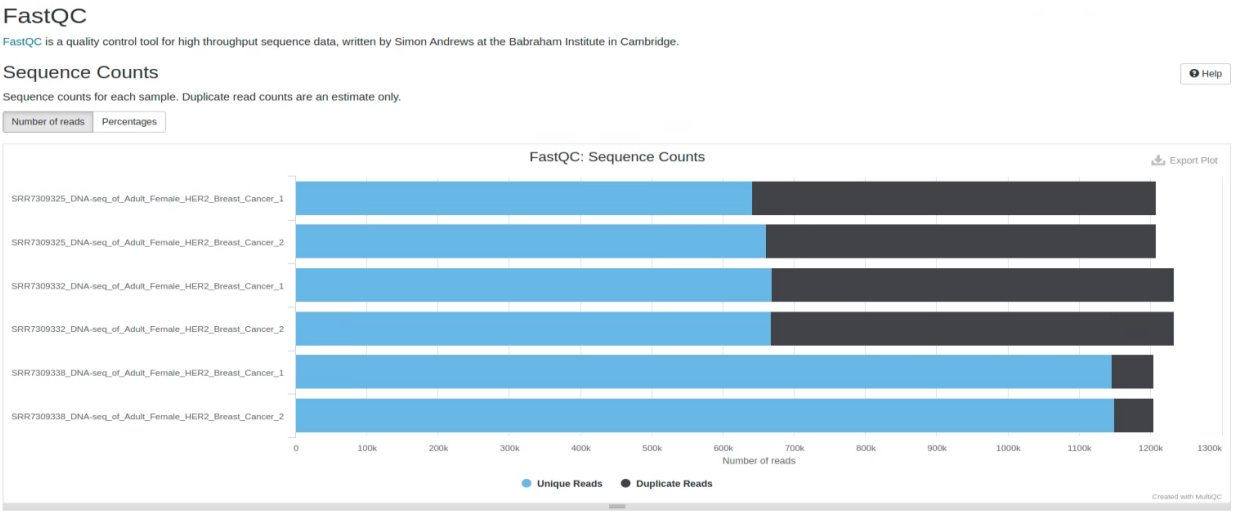


Figure 4 : Sequence Counts .Sequence counts for each sample. Duplicate read counts are an estimate only.

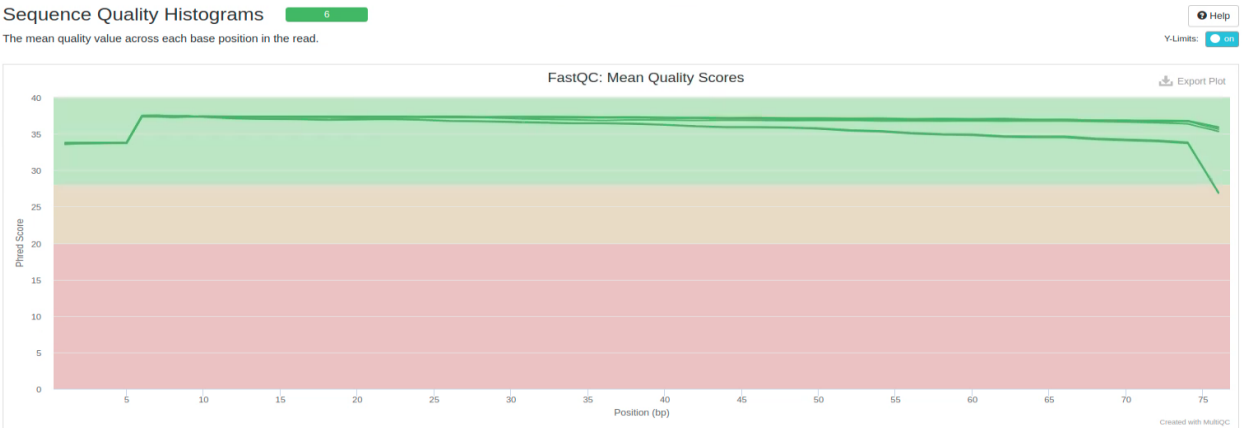


Figure 5 : Sequence Quality Histograms ．The mean quality value across each base position in the read.

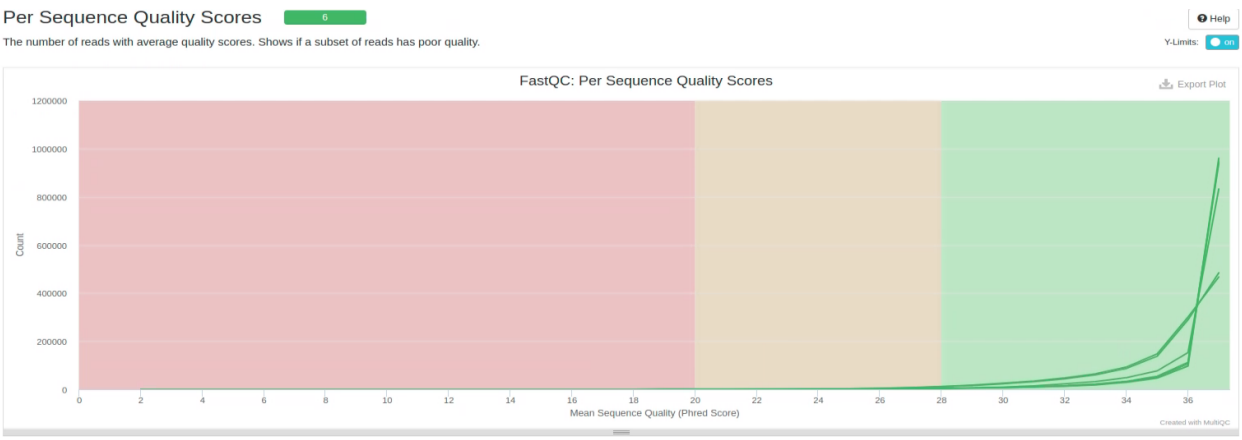


Figure 6 :Per Sequence Quality Scores .The number of reads with average quality scores. Shows if a subset of reads has poor quality.

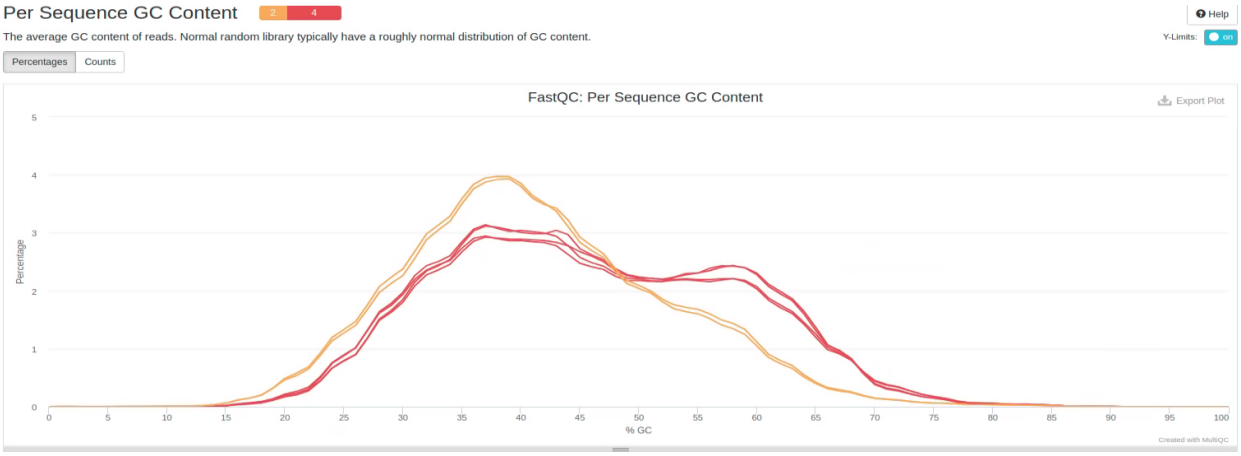


Figure 7 :Per Sequence GC Content.The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content.

**2- Alignment with BWA.**

**We took an overview about different pipelines and aligners from this paper (**Performance Assessment of Variant Calling Pipelines using Human Whole Exome Sequencing) to choose BWA aligner (https://www.biorxiv.org/content/10.1101/359109v1.full.pdf).BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences ranged from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

For all the algorithms, BWA first needs to construct the FM-index for the reference genome (the **index** command). Alignment algorithms are invoked with different sub-commands: **aln**/**samse**/**sampe** for BWA-backtrack, **bwasw** for BWA-SW and **mem** for the BWA-MEM algorithm.

**Alignment results :**

For the first sample(SRR7309332), there were 50.35% of reads mapped and only 46.69% properly mapped. However, for SRR7309338, there were 58.49% of reads mapped and only 47.57% properly mapped. The last one, SRR7309338, there were 46.55% of reads mapped and only 42.69% properly mapped.

```
2422464 + 0 in total (QC-passed reads + QC-failed reads)
6684 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
1127840 + 0 mapped (46.56% : N/A)
2415780 + 0 paired in sequencing
1207890 + 0 read1
1207890 + 0 read2
1029268 + 0 properly paired (42.61% : N/A)
1078442 + 0 with itself and mate mapped
42714 + 0 singletons (1.77% : N/A)
29328 + 0 with mate mapped to a different chr
14884 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 8 : SRR7309325_L001_R1.fastq.gz.sam_stats.out

```
2473199 + 0 in total (QC-passed reads + QC-failed reads)
6117 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
1245182 + 0 mapped (50.35% : N/A)
2467082 + 0 paired in sequencing
1233541 + 0 read1
1233541 + 0 read2
1151924 + 0 properly paired (46.69% : N/A)
1197350 + 0 with itself and mate mapped
41715 + 0 singletons (1.69% : N/A)
26688 + 0 with mate mapped to a different chr
15368 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 9 : SRR7309332_L002_R1.fastq.gz.sam_stats.out

```
2426519 + 0 in total (QC-passed reads + QC-failed reads)
16493 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
1419391 + 0 mapped (58.49% : N/A)
2410026 + 0 paired in sequencing
1205013 + 0 read1
1205013 + 0 read2
1146468 + 0 properly paired (47.57% : N/A)
1344574 + 0 with itself and mate mapped
58324 + 0 singletons (2.42% : N/A)
135464 + 0 with mate mapped to a different chr
43884 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 10 : SRR7309338_L002_R1.fastq.gz.sam_stats.out

### 3- **Duplicate reads were marked by Picard tools**

This tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA. Duplicates can arise during sample preparation e.g. library construction using PCR. See also EstimateLibraryComplexity for additional notes on PCR duplication artifacts. Duplicate reads can also result from a single amplification cluster, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument. These duplication artifacts are referred to as optical duplicates.

The MarkDuplicates tool works by comparing sequences in the 5 prime positions of both reads and read-pairs in a SAM/BAM file. An BARCODE_TAG option is available to facilitate duplicate marking using molecular barcodes. After duplicate reads are collected, the tool differentiates the primary and duplicate reads using an algorithm that ranks reads by the sums of their base-quality scores (default method).

The tool's main output is a new SAM or BAM file, in which duplicates have been identified in the SAM flags field for each read. Duplicates are marked with the hexadecimal value of 0x0400, which corresponds to a decimal value of 1024. If you are not familiar with this type of annotation.

Although the bitwise flag annotation indicates whether a read was marked as a duplicate, it does not identify the type of duplicate. To do this, a new tag called the duplicate type (DT) tag was recently added as an optional output in the 'optional field' section of a SAM/BAM file. Invoking the TAGGING_POLICY option, you can instruct the program to mark all the duplicates (All), only the optical duplicates (OpticalOnly), or no

duplicates (DontTag). The records within the output of a SAM/BAM file will have values for the 'DT' tag (depending on the invoked TAGGING_POLICY), as either library/PCR-generated duplicates (LB), or sequencing-platform artifact duplicates (SQ). This tool uses the READ_NAME_REGEX and the OPTICAL_DUPLICATE_PIXEL_DISTANCE options as the primary methods to identify and differentiate duplicate types. Set READ_NAME_REGEX to null to skip optical duplicate detection, e.g. for RNA-seq or other data where duplicate sets are extremely large and estimating library complexity is not an aim. Note that without optical duplicate counts, library size estimation will be inaccurate.

MarkDuplicates also produces a metrics file indicating the numbers of duplicates for both single- and paired-end reads.

The program can take either coordinate-sorted or query-sorted inputs, however the behavior is slightly different. When the input is coordinate-sorted, unmapped mates of mapped records and supplementary/secondary alignments are not marked as duplicates. However, when the input is query-sorted (actually query-grouped), then unmapped mates and secondary/supplementary reads are not excluded from the duplication test and can be marked as duplicate reads.

```
2422464 + 0 in total (QC-passed reads + QC-failed reads)
6684 + 0 secondary
0 + 0 supplementary
268479 + 0 duplicates
1127840 + 0 mapped (46.56% : N/A)
2415780 + 0 paired in sequencing
1207890 + 0 read1
1207890 + 0 read2
1029268 + 0 properly paired (42.61% : N/A)
1078442 + 0 with itself and mate mapped
42714 + 0 singletons (1.77% : N/A)
29328 + 0 with mate mapped to a different chr
14884 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 11: SRR7309325_L001_R1.fastq.gz.dedup.stat

```
2473199 + 0 in total (QC-passed reads + QC-failed reads)
6117 + 0 secondary
0 + 0 supplementary
265353 + 0 duplicates
1245182 + 0 mapped (50.35% : N/A)
2467082 + 0 paired in sequencing
1233541 + 0 read1
1233541 + 0 read2
1151924 + 0 properly paired (46.69% : N/A)
1197350 + 0 with itself and mate mapped
41715 + 0 singletons (1.69% : N/A)
26688 + 0 with mate mapped to a different chr
15368 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 12 : SRR7309332_L002_R1.fastq.gz.dedup.stat

```
2426519 + 0 in total (QC-passed reads + QC-failed reads)
16493 + 0 secondary
0 + 0 supplementary
38802 + 0 duplicates
1419391 + 0 mapped (58.49% : N/A)
2410026 + 0 paired in sequencing
1205013 + 0 read1
1205013 + 0 read2
1146468 + 0 properly paired (47.57% : N/A)
1344574 + 0 with itself and mate mapped
58324 + 0 singletons (2.42% : N/A)
135464 + 0 with mate mapped to a different chr
43884 + 0 with mate mapped to a different chr (mapQ>=5)
```

Figure 13 : SRR7309338_L002_R1.fastq.gz.dedup.stat

**4- local realignment and base recalibration will be conducted with GATK**

The GATK is the industry standard for identifying SNPs and indels in germline DNA and RNAseq data. Its scope is now expanding to include somatic short variant calling, and to tackle copy number (CNV) and structural variation (SV). In addition to the variant callers themselves, the GATK also includes many utilities to perform related tasks such as processing and quality control of high-throughput sequencing data, and bundles the popular Picard toolkit.

These tools were primarily designed to process exomes and whole genomes generated with Illumina sequencing technology, but they can be adapted to handle a variety of other technologies and experimental designs. And although it was originally developed for human genetics, the GATK has since evolved to handle genome data from any organism, with any level of ploidy.

Base quality scores are per-base estimates of error emitted by the sequencing machines; they express how confident the machine was that it called the correct base each time. For example, let's say the machine reads an A nucleotide, and assigns a quality score of Q20 -- in Phred-scale, that means it's 99% sure it identified the base correctly. This may seem high, but it does mean that we can expect it to be wrong in one case out of 100; so if we have several billion base calls (we get ~90 billion in a 30x genome), at that rate the machine would make the wrong call in 900 million bases -- which is a lot of bad bases. The quality score each base call gets is determined through some dark magic jealously guarded by the manufacturer of the sequencing machines.

Why does it matter? Because our short variant calling algorithms rely heavily on the quality score assigned to the individual base calls in each sequence read. This is because the quality score tells us how much we can trust that particular observation to inform us about the biological truth of the site where that base aligns. If we have a base call that has a low quality score, that means we're not sure we actually read that A correctly, and it could actually be something else. So we won't trust it as much as other base calls that have higher qualities. In other words we use that score to weigh the evidence that we have for or against a variant allele existing at a particular site.

Unfortunately the scores produced by the machines are subject to various sources of systematic (non-random) technical error, leading to over- or under-estimated base quality scores in the data. Some of these errors are due to the physics or the chemistry of how the sequencing reaction works, and some are probably due to manufacturing flaws in the equipment.

Base quality score recalibration (BQSR) is a process in which we apply machine learning to model these errors empirically and adjust the quality scores accordingly. For example we can identify that, for a given run, whenever we called two A nucleotides in a row, the next base we called had a 1% higher rate of error. So any base call that comes after AA in a read should have its quality score reduced by 1%. We do that over several different covariates (mainly sequence context and position in read, or cycle) in a way that is additive. So the same base may have its quality score increased for one reason and decreased for another.

This allows us to get more accurate base qualities overall, which in turn improves the accuracy of our variant calls. To be clear, we can't correct the base calls themselves, *i.e.* we can't determine whether that low-quality A should actually have been a T -- but we can at least tell the variant caller more accurately how far it can trust that A. Note that in some cases we may find that some bases should have a higher quality score, which

allows us to rescue observations that otherwise may have been given less consideration than they deserve. Anecdotally our impression is that sequencers are more often over-confident than under-confident, but we do occasionally see runs from sequencers that seemed to suffer from low self-esteem.

This procedure can be applied to BAM files containing data from any sequencing platform that outputs base quality scores on the expected scale. We have run it ourselves on data from several generations of Illumina, SOLiD, 454, Complete Genomics, and Pacific Biosciences sequencers.

The base recalibration process involves two key steps: first the BaseRecalibrator tool builds a model of covariation based on the input data and a set of known variants, producing a recalibration file; then the ApplyBQSR tool adjusts the base quality scores in the data based on the model, producing a new BAM file. The known variants are used to mask out bases at sites of real (expected) variation, to avoid counting real variants as errors. Outside of the masked sites, every mismatch is counted as an error. The rest is mostly accounting.

**5- Joint variant calling which includes many steps:**

A) assess genotype likelihood per-sample using HaplotypeCaller
B) combine samples using CombineGVCFs
C) Joint Genotyping using GenotypeGVCFs

Call germline SNPs and indels via local re-assembly of haplotypes

The HaplotypeCaller is capable of calling SNPs and indels simultaneously via local de-novo assembly of haplotypes in an active region. In other words, whenever the program encounters a region showing signs of variation, it discards the existing mapping information and completely reassembles the reads in that region. This allows the HaplotypeCaller to be more accurate when calling regions that are traditionally difficult to call, for example when they contain different types of variants close to each other. It also makes the HaplotypeCaller much better at calling indels than position-based callers like UnifiedGenotyper.

In the GVCF workflow used for scalable variant calling in DNA sequence data, HaplotypeCaller runs per-sample to generate an intermediate GVCF (not to be used in final analysis), which can then be used in GenotypeGVCFs for joint genotyping of multiple samples in a very efficient way. The GVCF workflow enables rapid incremental processing of samples as they roll off the sequencer, as well as scaling to very large cohort sizes (e.g. the 92K exomes of ExAC).

In addition, HaplotypeCaller is able to handle non-diploid organisms as well as pooled experiment data. Note however that the algorithms used to calculate variant likelihoods is not well suited to extreme allele frequencies (relative to ploidy) so its use is not recommended for somatic (cancer) variant discovery. For that purpose, use Mutect2 instead.

Finally, HaplotypeCaller is also able to correctly handle the splice junctions that make RNAseq a challenge for most variant callers.

## How HaplotypeCaller works

### 1. Define active regions

The program determines which regions of the genome it needs to operate on (active regions), based on the presence of evidence for variation.

### 2. Determine haplotypes by assembly of the active region

For each active region, the program builds a De Bruijn-like graph to reassemble the active region and identifies what are the possible haplotypes present in the data. The program then realigns each haplotype against the reference haplotype using the Smith-Waterman algorithm in order to identify potentially variant sites.

### 3. Determine likelihoods of the haplotypes given the read data

For each active region, the program performs a pairwise alignment of each read against each haplotype using the PairHMM algorithm. This produces a matrix of likelihoods of haplotypes given the read data. These likelihoods are then marginalized to obtain the likelihoods of alleles for each potentially variant site given the read data.

### 4. Assign sample genotypes

For each potentially variant site, the program applies Bayes' rule, using the likelihoods of alleles given the read data to calculate the likelihoods of each genotype per sample given the read data observed for that sample. The most likely genotype is then assigned to the sample.

## Input

Input bam file(s) from which to make variant calls

## Output

Either a VCF or GVCF file with raw, unfiltered SNP and indel calls. Regular VCFs must be filtered either by variant recalibration (Best Practice) or hard-filtering before use in downstream analyses. If using the GVCF workflow, the output is a GVCF file that must first be run through GenotypeGVCFs and then filtering before further analysis.

variant got annotated=9048

```
Location                      : raw_variants_ann.vcf.gz
Failed Filters                : 0
Passed Filters                : 34121

Sample Name: SRR7309325
SNPs                          : 6501
MNPs                          : 0
Insertions                    : 915
Deletions                     : 431
Indels                        : 0
Same as reference             : 1052
Missing Genotype              : 25208
Partial Genotype              : 14
Phased Genotypes              : 65.3% (5819/8913)
SNP Transitions/Transversions: 1.86 (7406/3990)
Total Het/Hom ratio           : 0.33 (1962/5899)
SNP Het/Hom ratio             : 0.33 (1612/4889)
MNP Het/Hom ratio             : - (0/0)
Insertion Het/Hom ratio       : 0.23 (169/746)
Deletion Het/Hom ratio        : 0.65 (170/261)
Indel Het/Hom ratio           : - (0/0)
Insertion/Deletion ratio      : 2.12 (915/431)
Indel/SNP+MNP ratio           : 0.21 (1346/6501)
```

Figure 14: SRR7309325_L001_R1 VCF statistics

```
Sample Name: SRR7309332
SNPs                          : 6026
MNPs                          : 0
Insertions                    : 856
Deletions                     : 392
Indels                        : 0
Same as reference             : 1364
Missing Genotype              : 25462
Partial Genotype              : 21
Phased Genotypes              : 62.4% (5406/8659)
SNP Transitions/Transversions: 1.71 (6748/3937)
Total Het/Hom ratio           : 0.29 (1661/5634)
SNP Het/Hom ratio             : 0.30 (1375/4651)
MNP Het/Hom ratio             : - (0/0)
Insertion Het/Hom ratio       : 0.19 (139/717)
Deletion Het/Hom ratio        : 0.54 (138/254)
Indel Het/Hom ratio           : - (0/0)
Insertion/Deletion ratio      : 2.18 (856/392)
Indel/SNP+MNP ratio           : 0.21 (1248/6026)
```

Figure 15: SRR7309332_L002_R1 VCF statistics

```
Sample Name: SRR7309338
SNPs                          : 20499
MNPs                          : 0
Insertions                    : 667
Deletions                     : 1023
Indels                        : 1
Same as reference             : 1634
Missing Genotype              : 10291
Partial Genotype              : 6
Phased Genotypes              : 72.7% (17330/23830)
SNP Transitions/Transversions: 1.45 (23254/15983)
Total Het/Hom ratio           : 0.09 (1896/20300)
SNP Het/Hom ratio             : 0.09 (1769/18730)
MNP Het/Hom ratio             : - (0/0)
Insertion Het/Hom ratio       : 0.08 (48/619)
Deletion Het/Hom ratio        : 0.08 (72/951)
Indel Het/Hom ratio           : - (1/0)
Insertion/Deletion ratio      : 0.65 (667/1023)
Indel/SNP+MNP ratio           : 0.08 (1691/20499)
```

Figure 16 :SRR7309338_L002_R1 VCF statistics

**6- Split SNPs and indels using (SelectVariants),Assess the different filters in both known and novel to decide the threshold for the filtration in each filter.**

the most common SNPs only

statistics on the passed SNPs vcf file

```
Location                       : raw_variants_ann_SNP_clean_all_passed.vcf.gz
Failed Filters                 : 0
Passed Filters                 : 20404

Sample Name: SRR7309325
SNPs                           : 4838
MNPs                           : 0
Insertions                     : 0
Deletions                      : 0
Indels                         : 0
Same as reference              : 587
Missing Genotype               : 14972
Partial Genotype               : 7
Phased Genotypes               : 63.3% (3439/5432)
SNP Transitions/Transversions: 1.98 (5574/2821)
Total Het/Hom ratio            : 0.36 (1289/3556)
SNP Het/Hom ratio              : 0.36 (1284/3554)
MNP Het/Hom ratio              : - (0/0)
Insertion Het/Hom ratio        : - (0/0)
Deletion Het/Hom ratio         : - (0/0)
Indel Het/Hom ratio            : - (0/0)
Insertion/Deletion ratio       : - (0/0)
Indel/SNP+MNP ratio            : 0.00 (0/4838)
```

Figure 17 : statistics on the passed SNPs vcf file for SRR7309325

```
Sample Name: SRR7309332
SNPs                          : 4517
MNPs                          : 0
Insertions                    : 0
Deletions                     : 0
Indels                        : 0
Same as reference             : 796
Missing Genotype              : 15086
Partial Genotype              : 5
Phased Genotypes              : 60.3% (3206/5318)
SNP Transitions/Transversions: 1.77 (5080/2877)
Total Het/Hom ratio           : 0.32 (1089/3433)
SNP Het/Hom ratio             : 0.32 (1084/3433)
MNP Het/Hom ratio             : - (0/0)
Insertion Het/Hom ratio       : - (0/0)
Deletion Het/Hom ratio        : - (0/0)
Indel Het/Hom ratio           : - (0/0)
Insertion/Deletion ratio      : - (0/0)
Indel/SNP+MNP ratio           : 0.00 (0/4517)
```

Figure 18 : statistics on the passed SNPs vcf file for SRR7309332

```
Sample Name: SRR7309338
SNPs                          : 13217
MNPs                          : 0
Insertions                    : 0
Deletions                     : 0
Indels                        : 0
Same as reference             : 892
Missing Genotype              : 6291
Partial Genotype              : 4
Phased Genotypes              : 69.5% (9810/14113)
SNP Transitions/Transversions: 1.51 (15248/10106)
Total Het/Hom ratio           : 0.09 (1088/12133)
SNP Het/Hom ratio             : 0.09 (1084/12133)
MNP Het/Hom ratio             : - (0/0)
Insertion Het/Hom ratio       : - (0/0)
Deletion Het/Hom ratio        : - (0/0)
Indel Het/Hom ratio           : - (0/0)
Insertion/Deletion ratio      : - (0/0)
Indel/SNP+MNP ratio           : 0.00 (0/13217)
```

Figure 19 : statistics on the passed SNPs vcf file for SRR7309338

**7-extract known snps for further analysis in a text file = 6902**

```
rs201284905
rs60785419
rs1002525272
rs7580986
rs6709553
rs12714345
rs895453722
rs1324556715
rs10190980
rs1219385337
rs823241
rs823196
rs13035709
rs4853768
rs2385436
rs7424791
rs1456836086
rs2312926
rs10172165
rs1026356064
rs902753410
rs2314697
rs1356945355
rs4436933
rs1453777
rs12470380
rs9287696
rs10929518
rs792103
rs557028922
```

Figure 20 : List of SNPs

# Future Work

Make functional annotation.Genome annotation is the process of identifying functional elements along the sequence of a genome, thus giving meaning to it. It is necessary because the sequencing of DNA produces sequences of unknown function.

# References

1. GLOBOCAN. Breast Cancer estimated Incidence, mortality

and prevalence Worldwide in 2012.

2. Hennessy BT, Smith DL, Ram PT, Lu Y, Mills GB. Exploiting the PI3K/AKT pathway for cancer drug discovery. Nat Rev Drug Discov. 2005; 4:988–1004. doi: 10.1038/nrd1902.

3. Slamon DJ, Leyland-Jones B, Shak S, Fuchs H, Paton V, Bajamonde A, Fleming T, Eiermann W, Wolter J, Pegram M, Baselga J, Norton L. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med. 2001; 344:783–92. doi: 10.1056/NEJM200103153441101.

4. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, Gianni L, Baselga J, Bell R, Jackisch C, Cameron D, Dowsett M, Barrios CH, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med. 2005; 353:1659–72. doi: 10.1056/NEJMoa052306.

5. Elster N, Collins DM, Toomey S, Crown J, Eustace AJ, Hennessy BT. HER2-family signalling mechanisms, clinical implications and targeting in breast cancer. Breast Cancer

Res Treat. 2015; 149:5–15. doi: 10.1007/s10549-014-3250-x.

6. Fung C, Zhou P, Joyce S, Trent K, Yuan JM, Grandis JR, Weissfeld JL, Romkes M, Weeks DE, Egloff AM. Identification of epidermal growth factor receptor (EGFR) genetic variants that modify risk for head and neck squamous cell carcinoma. Cancer Lett. 2015; 357:549–56. doi: 10.1016/j.canlet.2014.12.008.

7. Alaoui-Jamali MA, Morand GB, da Silva SD. ErbB polymorphisms: insights and implications for response to targeted cancer therapeutics. Front Genet. 2015; 6:17. doi: 10.3389/fgene.2015.00017.