ITX Integrative Transcriptomics
Dr. K. Nieselt, Faculty of Science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

EyeTN workshop **March 30 - April 1 2015**

Genotyping

During this practical session you will learn how you call SNPs from an aligned *bam* file. Furthermore you will learn how you can evaluate these SNPs.

SNPs can be called with the software *GATK* (Genome Analysis ToolKit). However the first call is very time consuming on the whole human genome (approx. 6 hours). This is why we already calculated the first *vcf* file using the following command:

```
gatk -T UnifiedGenotyper -R data/2_mapping_DNA/hg19_complete.fasta
-I mapped.bam -o seitz/3_variantCalling/genotyping
```

In order for this command to work, a dictionary has to be created for the index. This can be done with the program *CreateSequenceDictionary* from the *picard-tools*. An example call would be:

```
CreateSequenceDictionary.jar R=ref.fasta O=ref.dict
```

Another important part is that the *bam* file has at least one read group specified (in the header with "@RG". This can also be added using the *picard tools*:

```
picard AddOrReplaceReadGroups I=input_bam O=output_bam RGLB=lib RGPL=illumina RGPU=4410
RGSM=Project
```

where RGLB is the Library, RGPL platform, RGPU the platform unit and RGSM the sample name.

Your task now is to filter this file for relevant SNPs and analyze them. You then will compare these SNPs to known SNPs in the SNPdb using IGV

The data for this practical session is stored in the "3_variantCalling" folder.

1. **Filter the *vcf* file**

   Filter the *vcf* file so that only position with a coverage of at least 10 are kept in the new file. (`vcftools`)

2. **Annovar**

   - Now you will try to find the effect of the filtered SNPs using *annovar*
   - The program is located in a subfolder called "annovar" in the data folder of this practical session. (`table_annovar.pl`)
   - use your filtered *vcf* file
   - the database is located in the folder "humandb"
   - we are using the genome build "hg19"
   - write the output into your folder

- make sure that annovar deletes all temporary files
- Use the following protocols:
  - refGene
  - cytoBand
  - genomicSuperDups
  - esp6500siv2_all
  - 1000g2014oct_all
  - 1000g2014oct_afr
  - 1000g2014oct_eas
  - 1000g2014oct_eur
  - snp138
  - ljb26_all
- use the following operations: g,r,r,f,f,f,f,f,f,f

3. **Analyzing the data using IGV**

- Load the hg19 reference genome from the server
- laod the dbSNP 1.3.7 from the server
- load the bam file from the data folder
- load the bed file from the data folder
- load the filtered vcf file from your folder