Practical Session 6

GotCloud Variant Calling & Samtools

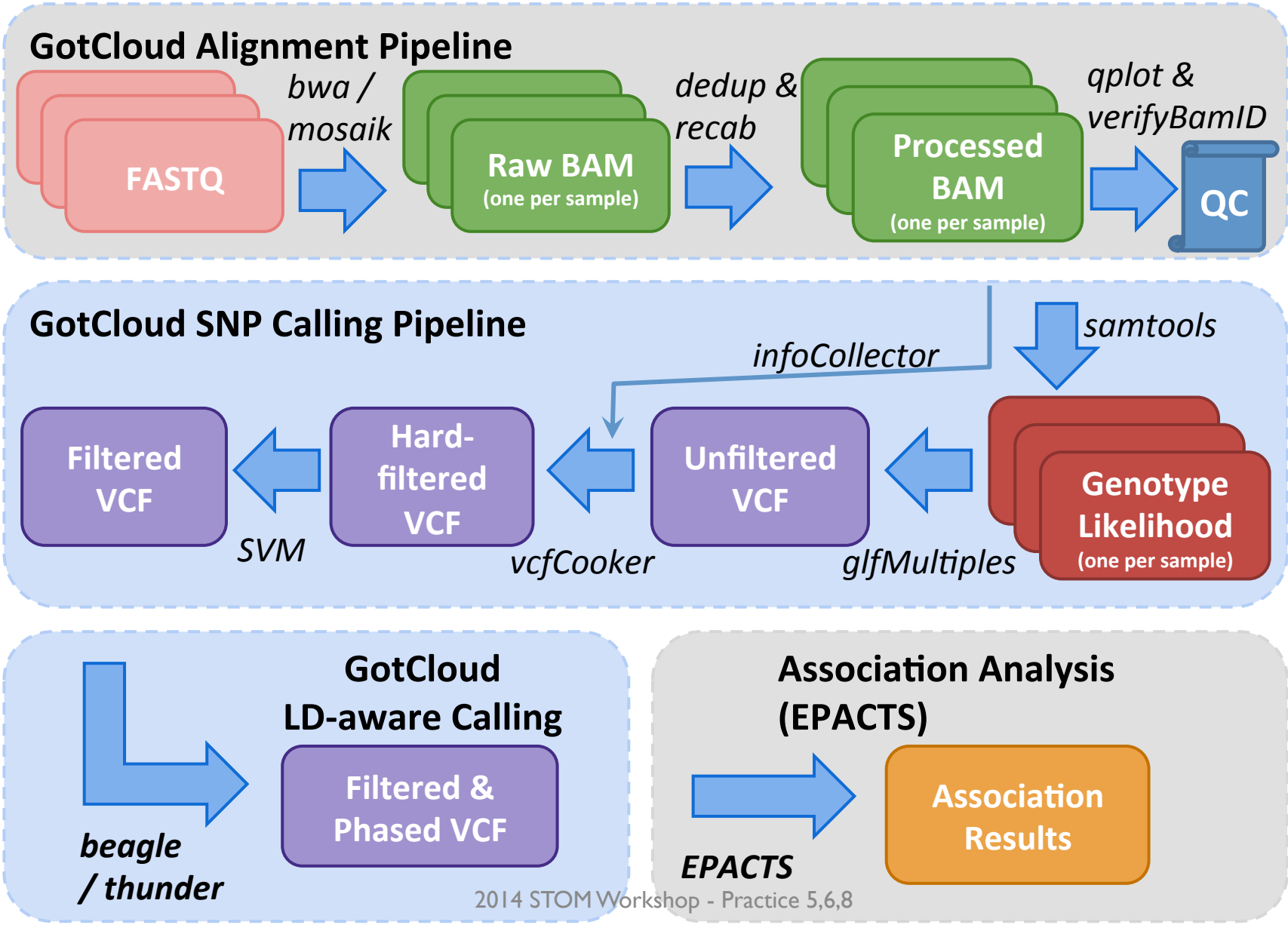
Jan 7th, 2014

STOM 2014 Workshop

Hyun Min Kang

University of Michigan, Ann Arbor

STEP 2 : GOTCLOUD SNP CALLING PIPELINE



INPUT DATA : LIST OF BAM FILES

```
% ls $S5/examples/bams
NA06984.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA06984.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam.bai
NA06985.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA06985.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam.bai
NA06986.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
NA06986.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam.bai
NA06989.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA06989.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam.bai
NA06994.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA06994.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam.bai
...
```

- The current example contains 99 CEU individuals in 500kb near CFTR gene

SNP CALLING : PREPARING INDEX FILE

```
% less $S5/examples/index/chr7.CFTR.low_coverage.index
```

```
NA06984 EUR     bams/NA06984.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA06985 EUR     bams/NA06985.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA06986 EUR     bams/NA06986.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
NA06989 EUR     bams/NA06989.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA06994 EUR     bams/NA06994.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA07000 EUR     bams/NA07000.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
NA07037 EUR     bams/NA07037.mapped.ILLUMINA.bwa.CEU.low_coverage.20130502.CFTR.bam
NA07048 EUR     bams/NA07048.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA07051 EUR     bams/NA07051.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA07056 EUR     bams/NA07056.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
NA07347 EUR     bams/NA07347.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
NA07357 EUR     bams/NA07357.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
NA10847 EUR     bams/NA10847.mapped.ILLUMINA.bwa.CEU.low_coverage.20130502.CFTR.bam
NA10851 EUR     bams/NA10851.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
NA11829 EUR     bams/NA11829.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
NA11830 EUR     bams/NA11830.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA11831 EUR     bams/NA11831.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA11832 EUR     bams/NA11832.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA11840 EUR     bams/NA11840.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA11843 EUR     bams/NA11843.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA11881 EUR     bams/NA11881.mapped.ILLUMINA.bwa.CEU.low_coverage.20120522.CFTR.bam
NA11892 EUR     bams/NA11892.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam
```

SNP CALLING : PREPARING CONFIGURATION FILE

```
% cat $S5/examples/index/chr7.CFTR.low_coverage.conf
CHRS = 7
BAM_INDEX = index/chr7.CFTR.low_coverage.index
############
# References
REF_ROOT = chr7Ref
#
REF = $(REF_ROOT)/hs37d5.chr7.fa
INDEL_PREFIX = $(REF_ROOT)/1kg.pilot_release.merged.indels.sites.hg19
DBSNP_VCF =  $(REF_ROOT)/dbsnp_135.b37.chr7.CFTR.vcf.gz
HM3_VCF =  $(REF_ROOT)/hapmap_3.3.b37.sites.chr7.CFTR.vcf.gz
OMNI_VCF = $(REF_ROOT)/1000G_omni2.5.b37.sites.PASS.chr7.CFTR.vcf.gz
```

RUNNING SNP CALLING

```
% time $S5/gotcloud/gotcloud snpcall --conf
  $S5/examples/index/chr7.CFTR.low_coverage.conf
  --outDir ~/out/snps --baseprefix $S5/examples
  --region 7:117000000-117500000 --numjobs 4
```

```
Key configurations:
GOTCLOUD_ROOT: /data/stom2014/session5/gotcloud
OUT_DIR:       /home/hmkang/out/snps
BAM_INDEX:     index/chr7.CFTR.low_coverage.index
REF:           chr7Ref/hs37d5.chr7.fa
CHRS:          7
BATCH_TYPE:    local
BATCH_OPTS:

Processing the following steps...
2: RUN_PILEUP
3: RUN_GLFMULTIPLES
4: RUN_VCFPILEUP
5: RUN_FILTER
6: RUN_SVM
8: RUN_SPLIT
Call region is 7:117000000-117500000
Generating commands for chr7...
Creating glf INDEX at 7:115000001-120000000..
```

Will take ~5 mins to finish

SUMMARY STATISTICS FROM VARIANT CALLS

```
% cat ~/out/snps/vcfs/chr7/chr7.filtered.sites.vcf.summary
```

FILTER	#SNPs	#dbSNP	%dbSNP	%CpG Known	%CpG Novel	%Known Ts/Tv	%Novel Ts/Tv	%nCpG-K Ts/Tv	%nCpG-N Ts/Tv	%HM3 sens	%HM3 /SNP
INDEL5	11	8	72.7	25.0	0.0	3.00	0.50	2.00	0.50	0.000	0.000
INDEL5;SVM	8	4	50.0	50.0	25.0	0.33	0.33	1.00	0.00	0.000	0.000
PASS	1388	1186	85.4	15.5	11.4	2.10	1.62	1.79	1.42	85.083	11.095
SVM	131	33	25.2	9.1	9.2	1.06	0.85	0.88	0.78	0.000	0.000

FILTER	#SNPs	#dbSNP	%dbSNP	%CpG Known	%CpG Novel	%Known Ts/Tv	%Novel Ts/Tv	%nCpG-K Ts/Tv	%nCpG-N Ts/Tv	%HM3 sens	%HM3 /SNP
INDEL5	19	12	63.2	33.3	14.3	1.40	0.40	1.67	0.20	0.000	0.000
PASS	1388	1186	85.4	15.5	11.4	2.10	1.62	1.79	1.42	85.083	11.095
SVM	139	37	26.6	13.5	9.8	0.95	0.82	0.88	0.74	0.000	0.000

PASS	1388	1186	85.4	15.5	11.4	2.10	1.62	1.79	1.42	85.083	11.095
FAIL	150	45	30.0	15.6	9.5	1.14	0.81	1.00	0.73	0.000	0.000
TOTAL	1538	1231	80.0	15.5	10.7	2.05	1.27	1.75	1.12	85.083	10.013

LD-AWARE GENOTYPE REFINEMENT

```
% time $S5/gotcloud/gotcloud beagle --conf
  $S5/examples/index/chr7.CFTR.low_coverage.conf
  --outDir ~/out/snps --baseprefix $S5/examples
  --region 7:117000000-117500000 --numjobs 2
```

```
Key configurations:
GOTCLOUD_ROOT: /data/stom2014/session5/gotcloud
OUT_DIR:       /home/hmkang/out/snps
BAM_INDEX:     index/chr7.CFTR.low_coverage.index
REF:           chr7Ref/hs37d5.chr7.fa
CHRS:          7
BATCH_TYPE:    local
BATCH_OPTS:

Processing the following steps...
9: RUN_BEAGLE
10: RUN_SUBSET
Call region is 7:117000000-117500000
Generating commands for chr7...
------------------------------------------------------------
Finished creating makefile /home/hmkang/out/snps/umake.Makefile

Running /home/hmkang/out/snps/umake.Makefile
```

Takes
~2 mins
to finish

CHECKING THE OUTPUT VCF FILES

```
zless ~/out/snps/beagle/chr7/chr7.filtered.PASS.beagled.vcf.gz
```

```
##FILTER=<ID=mq0,Description="Mapping Quality Below 0">
##FILTER=<ID=dp1,Description="Total Read Depth Below 1">
##FILTER=<ID=DP10000000,Description="Total Read Depth Above 10000000">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Most Likely Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Call Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=PL,Number=3,Type=Integer,Description="Genotype Likelihoods for Genotypes 0/0,0/1,1/1">
##FORMAT=<ID=PL3,Number=6,Type=Integer,Description="Genotype Likelihoods for Genotypes 0/0,0/1,1/1,0/2,1/2,2/2">
##FORMAT=<ID=BD,Number=1,Type=Float,Description="Genotype dosage from beagle">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  NA06984 NA06985 NA06986 NA06989 NA06994 NA07000 NA07037
NA07048 NA07051 NA07056 NA07347 NA07357 NA10847 NA10851 NA11829 NA11830 NA11831 NA11832 NA11840 NA11843 NA11881 NA11892 NA11893
NA11894 NA11918 NA11919 NA11920 NA11930 NA11931 NA11932 NA11933 NA11992 NA11994 NA11995 NA12003 NA12004 NA12005 NA12006 NA12043
NA12044 NA12045 NA12046 NA12058 NA12144 NA12154 NA12155 NA12156 NA12234 NA12249 NA12272 NA12273 NA12275 NA12282 NA12283 NA12286
NA12287 NA12340 NA12341 NA12342 NA12347 NA12348 NA12383 NA12399 NA12400 NA12413 NA12414 NA12489 NA12546 NA12716 NA12717 NA12718
NA12748 NA12749 NA12750 NA12751 NA12760 NA12761 NA12762 NA12763 NA12775 NA12776 NA12777 NA12778 NA12812 NA12813 NA12814 NA12815
NA12827 NA12828 NA12829 NA12830 NA12842 NA12843 NA12872 NA12873 NA12874 NA12878 NA12889 NA12890
7       117000961       .       G       A       100     PASS    DP=760;MQ=59;NS=95;AN=198;AC=4;AF=0.016671;AB=0.4393;AZ=-0.6573;
FIC=-0.0178;SLRT=-0.0535;HWEAF=0.0166;HWDAF=0.0333,0.0000;LBS=0,0,0,0,9,7,0,0;OBS=12,3,0,0,361,377,0,0;STR=-0.084;STZ=-2.309;CBR
=0.053;CBZ=1.464;IOR=0.000;IOZ=-1.003;AOI=-18.339;AOZ=-17.336;LQR=0.021;MQ0=0.000;MQ10=0.000;MQ20=0.000;MQ30=0.017;SVM=0.800812;
BAVGPOST=0.992;BRSQ=0.836       GT:GD:GQ:PL:BD  0|0:10:45:0,30,212:0    0|0:23:84:0,69,255:0    0|0:11:48:0,33,248:0    0|0:5:30
:0,15,137:0.0002        0|0:9:42:0,27,175:0     0|0:7:36:0,21,210:0.0001        0|0:6:33:0,18,188:0.0001        0|0:7:36:0,21,18
4:0.0001        0|0:0:15:0,0,0:0.0076   0|0:1:18:0,3,38:0.0032  0|0:16:63:0,48,253:0    0|0:10:45:0,30,242:0    0|0:14:57:0,42,2
55:0    0|0:3:24:0,9,106:0.0012 0|0:11:48:0,33,223:0    1|0:0:15:0,0,0:0.405    0|0:1:18:0,3,32:0.0039  0|0:6:33:0,18,154:0.0001
        0|0:7:36:0,21,149:0.0002        0|0:5:30:0,15,158:0.0002        0|0:7:36:0,21,189:0.0001        0|0:4:27:0,12,115:0.0005
        0|0:3:24:0,9,79:0.0008  0|0:3:24:0,9,101:0.001  0|0:7:36:0,21,145:0.0001        0|0:11:48:0,33,250:0    0|0:8:39:0,24,18
0:0     0|0:10:45:0,30,208:0    0|0:3:24:0,9,86:0.0012  0|1:4:47:62,0,41:1      0|1:12:100:133,0,129:1  0|0:10:45:0,30,177:0
```

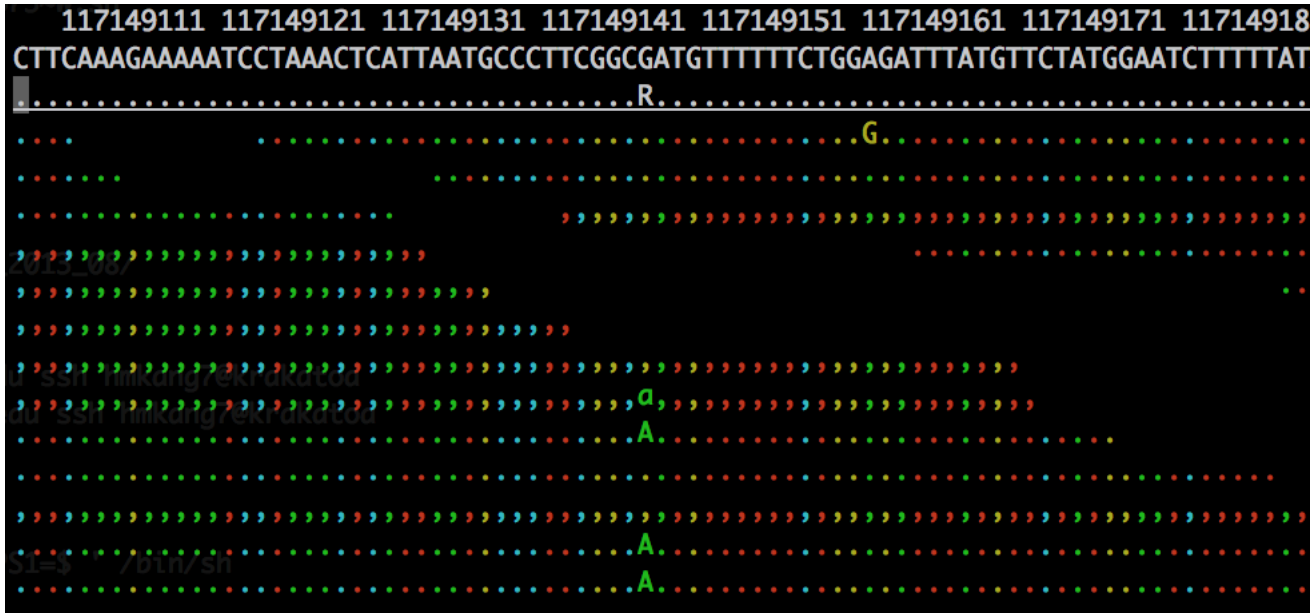
USING TABIX TO ACCESS REGION OF INTEREST

```
$S5/epacts/bin/tabix ~/out/snps/beagle/chr7/
chr7.filtered.PASS.beagled.vcf.gz 7:117149147-117149147
```

```
7       117149147       .       G       A       100     PASS    DP=812;MQ=59;NS=99;AN=198;AC=9;AF=0.046952;AB=0.5047;AZ
=0.0783;FIC=-0.0522;SLRT=-0.4591;HWEAF=0.0469;HWDAF=0.0941,0.0000;LBS=0,0,0,0,12,3,0,1;OBS=15,18,0,0,363,405,1,0;STR=0.
007;STZ=0.192;CBR=-0.029;CBZ=-0.808;IOR=0.938;IOZ=-0.064;AOI=-36.510;AOZ=-36.446;LQR=0.018;MQ0=0.000;MQ10=0.000;MQ20=0.
001;MQ30=0.017;SVM=1.94094;BAVGPOST=1.000;BRSQ=0.995       GT:GD:GQ:PL:BD  0|0:10:40:0,30,255:0    0|0:18:64:0,54,255:0
    0|0:18:64:0,54,255:0    0|1:4:10:20,0,64:1      0|0:10:40:0,30,180:0    0|0:10:40:0,30,252:0    0|0:10:40:0,30,248:
0    0|0:5:25:0,15,138:0.0003       0|0:1:13:0,3,32:0.0052  0|0:5:25:0,15,130:0.0002       0|0:15:55:0,45,255:0    0|
0:8:31:0,21,201:0.0002       0|0:9:37:0,27,222:0    0|0:3:19:0,9,98:0       0|0:7:31:0,21,192:0     0|0:2:16:0,6,58:0
     0|0:3:19:0,9,73:0.0011  0|0:16:58:0,48,255:0    0|0:10:40:0,30,204:0    0|0:8:34:0,24,222:0.0001       0|0:14:5
2:0,42,255:0    0|0:6:28:0,18,169:0.0002       0|0:10:40:0,30,238:0   0|0:6:28:0,18,179:0.0003 0|0:6:28:0,18,158:0.000
1       0|0:16:58:0,48,255:0    0|0:14:52:0,42,255:0    0|0:10:40:0,30,221:0   0|0:2:16:0,6,74:0       0|0:7:31:0,21,
189:0.0001      0|0:12:46:0,36,255:0    0|0:8:34:0,24,156:0     0|0:13:49:0,39,253:0   0|0:3:19:0,9,74:0       0|0:4
:22:0,12,82:0     0|0:14:52:0,42,249:0    0|0:15:52:0,42,255:0    0|0:14:52:0,42,250:0    0|0:9:37:0,27,209:0     0|0:
12:46:0,36,222:0    0|0:13:49:0,39,255:0    0|0:8:34:0,24,176:0     1|0:7:63:73,0,95:1      0|0:10:40:0,30,240:0    0|0
:8:34:0,24,164:0     0|0:19:67:0,57,255:0    0|0:3:19:0,9,70:0       0|0:10:40:0,30,247:0    0|0:4:22:0,12,141:0.0000|0
:3:19:0,9,96:0.0012   0|0:3:19:0,9,99:0.0097  0|0:4:22:0,12,99:0.0006 1|0:8:52:62,0,117:1     0|0:5:25:0,15,158:0.0003
     0|0:7:31:0,21,198:0     0|0:11:43:0,33,255:0    0|0:4:22:0,12,136:0.0003       0|0:8:31:0,21,123:0.0001
     0|0:18:64:0,54,255:0    0|0:8:34:0,24,190:0     0|0:8:34:0,24,224:0     0|0:15:55:0,45,255:0    0|0:8:34:0,24,1
61:0     0|0:5:25:0,15,102:0    0|1:5:41:51,0,66:1      0|0:7:31:0,21,152:0.0001       1|0:15:80:217,0,64:1    0|0:5:
25:0,15,136:0     1|0:8:72:147,0,56:1    0|0:4:22:0,12,136:0.0013 0|0:1:13:0,3,38:0.0084  0|0:10:40:0,30,248:0    0|0:1
0:40:0,30,255:0    0|0:14:52:0,42,255:0    0|0:10:40:0,30,247:0    0|0:8:34:0,24,174:0.0001       0|0:4:22:0,12,112:0
    0|0:7:31:0,21,168:0.0001       0|0:4:22:0,12,129:0.0000|0:5:25:0,15,108:0.0006 0|0:5:25:0,15,138:0.0003       0|
0:4:22:0,12,132:0     0|0:10:40:0,30,223:0    0|0:7:31:0,21,176:0.0000|0:2:16:0,6,66:0.0049   0|1:2:21:31,0,39:1      0
|0:7:31:0,21,197:0     0|0:5:25:0,15,155:0     0|0:4:22:0,12,105:0     0|0:14:52:0,42,255:0    0|0:7:31:0,21,155:0
0|0:12:46:0,36,255:0    0|1:10:97:107,0,127:1   0|0:11:43:0,33,255:0    0|0:4:22:0,12,123:0.0012       0|0:1:13:0,3,29
:0.0005  0|0:6:28:0,18,164:0.0003       0|1:9:45:55,0,145:1     0|0:11:43:0,33,240:0
```

Examining individual variant CFTR R75Q From BAM File Using Samtools

- `% samtools tview $S5/examples/bams/NA12843.mapped.ILLUMINA.bwa.CEU.low_coverage.20130415.CFTR.bam $S5/examples/chr7Ref/hs37d5.chr7.fa`
- Type ‘g’
- Type 7:117149147
- Type ‘n’ to color by nucleotide
- Move with left arrow to see the variant site clearly



SUMMARY : PRACTICAL SESSION 6

- If you have a list of BAM files already aligned
  - Can you call SNPs from the set of BAMs?
  - How can you check whether the quality of SNPs looks good or bad?
  - Can you refine the quality of genotypes by leveraging linkage disequilibrium structure?
  - Do you understand the format of VCF file?

- Can you examine the aligned sequence reads of a particular sample at a particular variant, to manually review whether the called SNPs looks good?