Dinh/Dinh 2012/NOTES/2012-11-13

From ZhangLabWiki

Contents

African methylomes - variant calling

- We need to (1) increase the size of the SNP matrix for mQTL and (2) compare the accuracy of SNP calls between different methods.
- Note: I handled the files using the index_id instead of sample_id. This greatly simplifies batch processing of those data files using shell scripts.
- Note: We do not have bam files from previous mapping, thus, new bam files were generated using an updated pipeline but still using SOAP2aligner.
- **BisRead** refers to our BisReadMapper pipeline and **BisSNP** refers to USC bisulfite methylation and SNP calling pipeline.

Increase the number of SNP calls

- I added in homozygous reference SNP calls to TPED file UPenn44.CGI-134.tped (very important for mQTL, less important for ASM).
  - The criteria for making a homozygous reference SNP calls is: depth >= 8 with base quality of Phred>5, and SNP Phred <= 5 (higher chance of being homozygous reference).
  - Reads were mapped using new methylation pipeline, and 1 bam file for each sample was generated (convert crick reads to watson, this is fine for most reads, but may need to ignore reads spanning indels. I assumed no indels).
- Shell script for extracting homozygous reference from **bam**:
  - Note: old version of samtools used.
  - Note: perl script used: File:ExtractHomoRefSNPs.txt

```
ref_fa="/media/2TB_storeA/BisRef/bisHg19/hg19.fa"
ref_fai="/media/2TB_storeA/BisRef/bisHg19/hg19.fa.fai"
samtools="/home/ddiep/softwares/samtools-0.1.8/samtools"
extractSNPs="./extractHomoRefSNPs.pl";
for g in `seq 1 1 48`
do
        f="Indx$g"
        echo "cd $f" > $f.job
        echo "$samtools pileup -Ac -l ../SNP_LIST -f $ref_fa $f.bam > $f.pileup" >> $f.job
        echo "$extractSNPs $f.pileup > $f.homoRef.snp" >> $f.job
        sh $f.job > $f.snp.log
done
```

- Example for Indx1.job:

```
cd Indx1
/home/ddiep/softwares/samtools-0.1.8/samtools pileup -Ac -l ../SNP_LIST -f /media/2TB_storeA/BisRef/bisHg19/hg19.fa Indx1.bam > Indx1.pileup
./extractHomoRefSNPs.pl Indx1.pileup > Indx1.homoRef.snp
```

- SNP calls were added to previous TPED/TFAM using: File:AddHomoRefToTPED.txt
  - Note: Must create TFAM_INDX file for this to work. Also, all Indx*homoRef.snp files must in the same directory with TPED/TFAM.
  - TFAM_INDX:

```
Indx19  CAFU028 0       0       0       -9
Indx18  CAFU042 0       0       0       -9
Indx17  CAFU043 0       0       0       -9
Indx16  CAMF013 0       0       0       -9
Indx20  CAMF022 0       0       0       -9
Indx1   CAPB016 0       0       0       -9
Indx21  CAPB043 0       0       0       -9
Indx2   CAPB046 0       0       0       -9
Indx3   CAPB056 0       0       0       -9
Indx24  CAPL036 0       0       0       -9
Indx22  CAPL049 0       0       0       -9
Indx4   CAPL056 0       0       0       -9
Indx23  CAPM001 0       0       0       -9
Indx36  CAPM003 0       0       0       -9
Indx37  CAPM004 0       0       0       -9
Indx5   CAPM007 0       0       0       -9
Indx39  ETAM042 0       0       0       -9
Indx42  ETAM058 0       0       0       -9
Indx43  ETAM065 0       0       0       -9
Indx40  ETAM071 0       0       0       -9
Indx41  ETAM077 0       0       0       -9
Indx44  ETSB008 0       0       0       -9
Indx45  ETSB027 0       0       0       -9
Indx46  ETSB031 0       0       0       -9
Indx47  ETSB035 0       0       0       -9
Indx48  ETSB036 0       0       0       -9
Indx25  KEBR007 0       0       0       -9
Indx27  KEBR042 0       0       0       -9
Indx29  KEBR061 0       0       0       -9
Indx30  KEPK003 0       0       0       -9
Indx31  KEPK006 0       0       0       -9
Indx32  KEPK007 0       0       0       -9
Indx33  KEPK010 0       0       0       -9
```

```
Indx34  KEPK016 0       0       0       -9
Indx6   TZHZ018 0       0       0       -9
Indx7   TZHZ033 0       0       0       -9
Indx8   TZHZ075 0       0       0       -9
Indx9   TZHZ214 0       0       0       -9
Indx10  TZHZ221 0       0       0       -9
Indx11  TZSW067 0       0       0       -9
Indx12  TZSW128 0       0       0       -9
Indx13  TZSW131 0       0       0       -9
Indx14  TZSW132 0       0       0       -9
Indx15  TZSW135 0       0       0       -9
```

- Finally, run plink to filter/clean. *I filtered out novel SNPs (no rs), because plink returned an error with more than 2 alleles found at those positions.

```
mv UPenn44.CGI-134.hRef.tped UPenn44.CGI-134.hRef.wNovel
grep -v chr: UPenn44.CGI-134.hRef.wNovel > UPenn44.CGI-134.hRef.tped
~ddiep/softwares/plink-1.07-x86_64/plink --tfile UPenn44.CGI-134.hRef --noweb --geno 0.25 --recode --transpose --out UPenn44.CGI-134.hRef.filtered
```

Make SNP calls using BisSNP (USC)

- BisSNP uses GATK based variant caller. BisSNP requires a reference dbSNP file in vcf format (provided on their website), and that only 1 bam file with crick positions mapped to watson is the input.
- To make BisSNP runs faster, we can give it a region file, so that it can ignore the majority of SNPs in dbSNP. I generated this file by taking out target regions (hg18), and used UCSC liftover tool to convert to hg19 coordinates. File:Hg19 regions miss24.txt * 24 target regions where not found in hg19.
- Shell script to run BisSNP:

```
ref_fa="/media/2TB_storeA/BisRef/bisHg19/hg19.fa"
ref_fai="/media/2TB_storeA/BisRef/bisHg19/hg19.fa.fai"
cpg_list="/media/2TB_storeA/BisRef/bisHg19/C_Pos/hg19.fa.cpg.positions.txt"
samtools="/home/ddiep/softwares/samtools-0.1.18/samtools"
tmp_dir="";

bisSnp="/home/ddiep/softwares/Bis-SNP/Utils/bissnp_easy_usage.pl --interval ../hg19_regions_miss24.bed /home/ddiep/softwares/Bis-SNP/BisSNP-0.71.jar"
vcf="/media/2TB_storeA/dbSNP/dbsnp_135.hg19.sort.vcf"


picards="java -Xmx4g -jar ~ddiep/softwares/picard-tools-1.74"

BASE_CHRS="chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 \
chr14 chr15 chr16 chr17 chr18 chr19 chr20 chr21 chr22 chrX chrY chrM"

for g in `seq 1 1 48`
do
        f="Indx$g"
        echo "cd $f" > $f.job
        echo "$samtools fillmd -b $f.bam $ref_fa > $f.fillmd.bam" >> $f.job
        echo "$picards/AddOrReplaceReadGroups.jar I=$f.fillmd.bam O=$f.rg.bam ID=HiSeq LB=HiSeq PL=illumina PU=HiSeq SM=$f CREATE_INDEX=true VALIDATION_STRINGENCY=SILENT" >> $f.job
        echo "$samtools index $f.rg.bam" >> $f.job
        echo "$bisSnp $f.rg.bam $ref_fa $vcf" >> $f.job

        nohup sh $f.job > BisSnp.$f.log
done
```

- Example Indx1.job:

```
cd Indx1
/home/ddiep/softwares/Bis-SNP/Utils/bissnp_easy_usage.pl --interval ../hg19_regions_miss24.bed /home/ddiep/softwares/Bis-SNP/BisSNP-0.71.jar Indx1.rg.bam /media/2TB_storeA/BisRef/bisHg19/hg
```

- After BisSNP finishes, move all *snp.raw.vcf files into one directory.
- Shell script for filtering out low quality SNP calls, converting VCF to TPED, and adding homozygous reference calls:

```
for f in `seq 1 1 48`
do
        grep -v LowQual Indx$f.rg.snp.raw.vcf > tmp.vcf
        ~/softwares/vcftools_0.1.9/bin/vcftools --vcf tmp.vcf --out Indx$f --plink-tped --recode
        awk '{ if($7 >= 8) print $1"\t"$2"\t0\t"$2"\t"$3"\t"$3}' ../HomoRefSNPs/Indx$f.homoRef.snp | sort -u > wHR.Indx$f.tped
        cat Indx$f.tped >> wHR.Indx$f.tped
        cp Indx$f.tfam wHR.Indx$f.tfam
done
```

Compare SNP calls accuracies

- We have 33 individuals with Illumina 1M duo SNPs calls (Hg18). TPED/TFAM: **PennAfrican_Batch1_genotypes**
- All SNPs called with BisSNP (USC) are on forward strand, but SNPs from the array could be on forward or reverse strand, thus needs to double check.
- Download reference file for array SNPs:

```
wget http://www.well.ox.ac.uk/~wrayner/strand/Human1M-Duov3_B-b36-strand.zip
```

- I had some issues before with some rs values in BisSNP calls being give more than 1 chromosome positions. To get the correct rs values, I used snp134_snv.txt (reduced from dbSNP134.txt)
- First, script to split TPED into individuals TPED. I wanted to rename the files with index_id, so I changed the sample_ids in TFAM to index_id for this step.

```
for f in `seq 1 1 33`
do
        head -n $f PennAfrican_Batch1_genotypes.tfam | tail -n 1 > keep.txt
        g=`awk '{print $2}' keep.txt`
        #echo "$f $g"
        ~ddiep/softwares/plink-1.07-x86_64/plink --noweb --tfile PennAfrican_Batch1_genotypes --keep keep.txt --recode --transpose --out PennAfrican_ArraySNP_$g
```

```
done
```

- Next, print genotypes from sequence and from array side by side:
  - Note: File:CompareWithArraySnps.txt

```
for f in PennAfrican_ArraySNP*tped
do
        g=`echo $f | sed 's/PennAfrican_ArraySNP_//g'`
        echo "wHR.$g"
        ./compareWithArraySnps.pl ../Latest_BisSNP_SNPs/BisSNP-wHR/wHR.$g $f > $g.compareSNPs
done
```

- Next, correct strand of array SNPs to match & count:
  - Note: File:CorrectStrand.txt

```
for f in *compareSNPs
do

        grep -v NA $f | grep -v 0:0 | ./correctStrand.pl > $f.Corrected
        total=`wc -l $f.Corrected`
        match=`awk '{if($4 == $5) print $0}' $f.Corrected | wc -l`
        echo $f $total $match
done
```

- Go back and check. The wrong Indx matches are around 50% while correct Indx matches are around 96%.
- SNPs printed in *compareSNPs are correct.
- For BisRead SNPs, split the TPED file into individual TPEDs as with for PennAfrican_Batch1_genotypes matrix, and compare TPED to TPED as with BisSNP.

Comparison of BisSNP and BisRead

- Conclusions:

```
(1)BisSNP gives 1-4% more SNPs than BisRead (using the method described above.)
(2)BisRead seems to be slightly more accurate than BisSNP when compared with Illumina 1M Duo array.
```

TFAM_ID	SAMPLE_ID	INDX_ID	#BisSNP_Compared	#BisSNP_Matched	%Matched	#BisRead_Compared	#BisRead_Matched	%Matched	#SNPs_bisSNP	#
577	CAPB016	Indx1	9803	9448	96%	9806	9533	97%	68856	6
355	TZSW067	Indx11	9386	9033	96%	9396	9101	97%	66681	6
606	TZSW128	Indx12	8729	8363	96%	8413	8116	96%	61422	5
262	TZSW131	Indx13	8868	8481	96%	8621	8360	97%	62949	6
309	TZSW135	Indx15	8878	8502	96%	8660	8389	97%	62831	6
200	CAMF013	Indx16	8980	8636	96%	8863	8610	97%	63539	6
79	CAFU043	Indx17	9582	9208	96%	9515	9256	97%	67271	6
184	CAFU042	Indx18	9062	8716	96%	8857	8586	97%	63564	6
376	CAFU028	Indx19	9036	8666	96%	8925	8650	97%	64176	6
470	CAPB046	Indx2	9525	9152	96%	9544	9278	97%	68074	6
158	CAMF022	Indx20	8951	8565	96%	8759	8469	97%	63215	6
735	CAPB043	Indx21	8904	8560	96%	8739	8483	97%	63290	6
604	CAPM001	Indx23	8868	8553	96%	8688	8451	97%	62896	6
498	CAPL036	Indx24	9172	8786	96%	9127	8846	97%	65118	6
742	KEBR007	Indx25	9349	8941	96%	9362	9049	97%	66511	6
729	KEBR028	Indx26	7829	7083	90%	0	0	NA	53277	N
110	KEBR042	Indx27	8965	8579	96%	8648	8346	97%	62888	6
762	KEBR061	Indx29	9182	8793	96%	9004	8718	97%	64874	6
705	CAPB056	Indx3	9327	8975	96%	9315	9046	97%	66169	6
750	KEPK003	Indx30	8862	8538	96%	8637	8351	97%	62578	6
732	KEPK006	Indx31	8735	8367	96%	8540	8256	97%	62054	6
718	KEPK007	Indx32	8652	8324	96%	8442	8175	97%	61132	5
749	KEPK010	Indx33	9663	9256	96%	9563	9276	97%	67529	6
743	KEPK016	Indx34	9199	8823	96%	8995	8743	97%	64606	6
651	CAPL056	Indx4	8967	8617	96%	8762	8481	97%	63478	6
716	ETSB008	Indx44	9281	8930	96%	9194	8924	97%	65210	6
788	ETSB027	Indx45	8839	8517	96%	8676	8396	97%	62471	6
717	ETSB031	Indx46	9080	8724	96%	8891	8635	97%	63680	6
719	ETSB035	Indx47	8993	8672	96%	8866	8615	97%	63165	6
759	ETSB036	Indx48	8971	8644	96%	8772	8516	97%	63517	6
728	TZHZ018	Indx6	9261	8936	96%	9263	9034	98%	66179	6
783	TZHZ075	Indx8	8792	8384	95%	8717	8323	95%	60491	5
463	TZHZ214	Indx9	9196	8803	96%	8881	8623	97%	63972	6

Retrieved from "http://genome-tech.ucsd.edu/LabNotes/index.php?title=Dinh/Dinh_2012/NOTES/2012-11-13&oldid=43796"

- This page was last modified on 14 November 2012, at 21:40.
-