BSNP: A BAYESIAN GENOTYPE CALLER

Brad Gulko – 18-May -2012 – V2.17.02a

Overview

The BSNP software[xix] was designed to derive unphased genotype polymorphism probabilities from short read alignments. Aligned short reads in `.SAM` or `.BAM` format[i] can be processed through the Samtools[ii], `pileup` utility[iii] to produce an output that serves as an input to BSNP. BSNP provides posterior genotype probabilities P(Genotype|Data) as well as full joint probability P(Genotype & Data), so the amount of data available to generate the posterior can be used in downstream analysis. BSNP uses a similar strategy to the one employed by SOAP[iv], but is open source, does not bias results toward a reference genome, and provides direct control of prior probability assignments. BSNP is coded in portable C++, and has no external dependencies so compiles easily under Linux or Windows. In limited testing BSBNP shows slightly better calling accuracy than MAQ[x] both for heterozygotes and homozygotes calls, when compared to HapMap[v] genotype calls.

Download

BSNP is presently available from the Siepel Lab[vi] website at http://compgen.bscb.cornell.edu/GPhoCS/BSNP (see Downloads table). If this address changes or becomes unavailable please feel free to contact the author (Brad Gulko) at the address provided in the section 5 History, below. The distribution is a compressed[vii] tar file commonly referred to as a tarball. It can be decompressed using WinRAR[viii] or 7Zip[ix] under Windows or the Linux commands:

```
bzcat BSNP_v2.17.02.tar.bz2 | tar -x
```

This should produce a top level directory containing Microsoft Visual Studio project files and a Linux `Makefile`, as well as the following subdirectories:

- `src` – contains all source and header files. No external dependencies.
- `bin` – for convenience, 32-bit x86 binaries are available here for Windows command lien and Linux access.
- `doc` – this document.
- `examples` – sample pileup file, along with sample outputs and instructions for regenerating the outputs.

Compilation

Under Linux, from the top level directory, type

```
make
```

This should regenerate an optimized binary file in the `bin` sub directory. There are neither external dependencies (except the standard libraries), nor any needed configuration. This code compiles without warnings or errors under relatively liberal compiler notifications with g++ (gcc) v4.1.2. Under Windows, either utilize the provide executables in `bin` or utilize Microsoft Visual Studioby opening the file `BSNP.sln`. in the top level directory, both optimized (release) and debugging profiles are provided. Other compilers under windows may be used and the code should compile with a command as simple as:

```
g++ *.cpp
```

or the equivalent executed from the `src` subdirectory.

BSNP Features:

- Provides complete posterior unphased genotype probability distribution over the 10 unphased diploid genotypes.
- Does not bias calculations towards a reference genome.
- Takes into account both read quality, and alignment quality on a per-nucleotide basis.
- Provides a flexible correlated observation correction configurable to mimic MAQ[x] or SOAP[xi] (can be deactivated).
- Provides P(Data) so full joint distribution P(Genotype & Data) can be reconstructed allowing for assessment of the amount of informative data driving the probability distribution and assisting in the principled Bayesian combination of results from different experiments.
- Provides a three parameter prior probability model to compactly account for most common priors.
- Both genotype priors may be entered directly for maximal control (in progress).
- Provides digest for observations that match the reference with high confidence.
- Provides summary information for positions not called as polymorphism candidates.
- Can provide sequence technology specific error model for: Illumina, 454, SOLiD and Sanger sequencing (experimental).
- All data need for calculations on a single position are in a single line of the input file, so it is relatively simple to parallelize software by splitting input file, running BSNP on split input files, then merging result files.
- Consistently more accurate for human genome calls (Compared to HapMap data) than the leading SNP caller, MAQ. The difference is small, but error rates are consistently lower, typically by about 3% of MAQ's error rate.

BSNP Limitations:

- Does not presently allow bias towards reference genome (simple update, contact author if you need this).
- Does not handle insertions or deletions, they are simply ignored.
- Defaults are convenient for conservative estimates for human genome and are based on statistics derived from hg18.
- BSNP was not optimized for speed. Running this on a complete human genome with coverage in the 20-40x range can take 44 hours on a single 2010 era CPU. However, memory load is typically light and generally O(n) in the read depth of the most deeply covered genomic position. As mentioned above, the process can easily be parallelized for multicore CPU's by splitting the input by chromosome, resulting in a linear speedup with the number of CPU's employed.
- Positions that have no short read coverage at all are generally not represented in the pileup file used as input to BSNP, and are neither represented in BSNP's SNP output nor its NonSNP output.

Contents

1 Primary Input File

BSNP currently accepts an ASCII summary file knows as a Pileup file. This file is generated by the Samtools pileup command, which takes as arguments aligned reads (generally in SAM / BAM file format, as produces by currently popular aligners such as BWA) and produces a summary file that lists the observed data (nucleotides, sequencing quality and alignment quality) at each genomic position. SNP can be easily modified to simply take in a set of nucleotides and produce a genotype probability distribution.

The `samtools pileup` command would be invoked as follows:

```
samtools pileup -csf hg18.fa Alignment.bam > Alignment.pileup
```

The hg18.fa is a human reference genome. While this reference is not used to calculate BSNP statistics, it is used to determine whether to send a given line of output to the genotype file (SNP) or the summary file (.nSNP) fasta file of the hg18 genome[xii].

Note: the `pileup` command is deprecated (2012) in Samtools and is being replaced by the `mpileup` command which is similar, but does not provide a simple MAQ call. The output formats are similar, but not quite the same. In a pinch, the MAQ call can be replaced with the reference and the subsequent MAQ related qualities can be spoofed. These will not affect the genotype probabilities generated by BSNP.

Sample Input File:

The pileup file is a tab delimited file containing fields as follow below[xiii]:

```
1     2     3     4     5     6     7     8     9
chr1  1254  t     W     27    41    22    35    a.,a,.gA..AA..,^.a,..^$,,^9A..^$..A$A$an.^2aAA,

                  10                                          11
                  8-8279)56565797-278347896998:!:+:::         :^94$$#7$.9/@+%.)$5$S944$0;8+9(2+W$
```

Field descriptions[xiv]:

1 Chromosome ID.
2 Position in chromosome (1 based, that is, the first position in a chromosome is 1, not 0)
3 Reference base
4 MAQ consensus base
5 Phred scaled quality of consensus base call
6 Phred scaled SNP quality (likelihood of a SNP, possible greater than likelihood of selected call)
7 Maximum mapping quality
8 Number of reads covering this position
9 List of actual reads covering this location where "." Means a match to the reference on the positive strand, "," means a match to the reference on the negative strand "^" indicates the start of a new read (skip the next character) and "$" means the end of a previous read. Capital characters mean reads on the forward strand (ACGTN), lower case characters represent reads on the reverse strand. For more details see the SAMTOOLS documentation on pileup or mpileup.
10 Read Qualities in phred scale (single ASCII[xv] character, where that character's ASCII value is $33 + 10 * \log_{10}(P)$, where P is the probability of an error)[xvi]
11 Alignment Qualities in phred scale.

2 INPUT ARGUMENTS:

```
# BSNP 2.15
# Usage:
BSNP  [-i S]  [-o S]  [-on S] [-os S] [-ph F] [-k F]  [-p0 F] [-ip S]
      [-mq I] [-mp F] [-th F] [-tm F] [-si B] [-ns B] [-ig B] [-v B]
      [-d B]  [-ti B] [-pb B]

           -i  : Input File              : - <- = stdin>
           -ip : Input Priors            : - <- = stdin
           -o  : Output File             : - <- = stdout>
           -on : Output File (NonSNPInfo): - <- = stderr>
           -os : Output File (SummryInfo): - <- = stdout>
           -ph : Phee                    : 0.500000
           -ka : Kappa                   : 1.000000
           -p0 : P0 (aka Pi, het Prob)   : 0.001000
           -ip : Genotype prior override :  <- <unused by default>
           -mq : Min MAQ SNP             : 1
           -mp : Min B SNP Probability   : 0.001000
           -th : Theta (0-1.0,1.0=off)   : 0.850000
           -tm : ThetaMin (0-1.0,0=off)  : 0.000000
           -si : Dump Summary Info       : 1
           -ns : Dump NonSNP Info        : 1
           -ig : Ignore Invalid Bases    : 1
           -v  : Verbose Operation       : 0
           -d  : Dump Debugging Info     : 0
           -ti : Invert Theta Sorting    : 0
           -pb : Print BSNP Call Char    : 1
           -st : Sequencing Technology   : 0
                      0 = UNKNOWN, 1 = Illumina\Solexa, 2 = Sanger,
                      3 = 454, 4 = SOLiD (converted to Nuc Space)
```

2.1 DESIGNATE INPUT FILES

`-i  : Input File              : - <- = stdin>`

As pileup files can be quite large (hundreds of gigabytes) it is often useful to keep them in a highly compressed format such as bzip2. The –i argument allows pileup files to be streamed out of a decompresser into the BSNP command, or read directly from a stored filesystem.

`-ip : Input Priors            : - <- = stdin>`

This allows the direct import of prior probabilities for each genotype. It overrides the 3 parameter model currently in place and automatically infers the individual (haploid) nucleotide priors from the set of 10 diploid priors provided in an input file by the user. The value for this argument is a string consisting of a file name, that file should contain a one line per genotype, with one line for each of the 10 diploid priors AA AC AG AT CC CG CT GG GT TT (case sensitive). Each line should contain {Geno Name Prob Endline}, for example AA 0.145\n. Lines beginning with a # character are ignored as comments.

2.2 DESIGNATE OUTPUT FILES

```
-o  : Output File              : - <- = stdout>
-on : Output File (NonSNPInfo): - <- = stderr>
-os : Output File (SummryInfo): - <- = stdout>
```

Three types of output files can be generated. For maximum flexibility, each can be directed to a stream thus large output files (SNP and nSNP) can be piped into a stream compressor for compact storage. However, there are some practical limitations:

- If the Summary Info is directed to stdout, than it will merge with the SNP output if that is also directed to stdout.
- If Verbose or Debugging modes are selected then this output will be merged with the stderr stream which may corrupt the nSNP data, if that data is also sent to stderr.

In general, it is recommended that the –os flag be used to direct summary output to a file (rather than stderr) and that either the .nSNP data be directed to a file, or that the Verbose and Debugging output be disabled. Under Linux, the bash named pipe operation ">(bzip2 > outfile.bz2)" can be used to replace a file name with a command the compressed the output stream and saves the compressed version. Beware, the created subshell may terminate asynchronously and much later than the parent process.

-o Output File SNP – genotype call file. It contains a line per position in the pileup file for which the genotype disagrees with the reference sequence, according to criteria provided in other options. This file can be large (>10Gb) and it may be useful to compress it via a pipe through bzip2 as it is being generated.

-on Output File nSNP – agreement summary file. Provides summary information for each position in the pileup file for which the genotype call agrees with the reference sequence, according to criteria provided in other options. This file can be large (>10Gb) and it may be useful to compress it via a pipe through bzip2 as it is being generated.

-os Output File Summary – provides useful summary statistics regarding the genome as a whole, such as number of genotype calls, number of nucleotides encountered, entropy etc.

2.3 PARAMETERS FOR PRIOR MODEL

```
-ph : Phee                     : 0.500000
-ka : Kappa                    : 1.000000
-p0 : P0 (aka Pi, het Prob)    : 0.001000
```

These three parameters are sufficient for defining a prior probability distribution over individual nucleotides and genotypes.

`-p0` represents the probability of a heterozygous genotype $p_0 = \frac{(p_{ac}+p_{ag}+p_{at}+p_{cg}+p_{ct}+p_{gt})}{(p_{aa}+p_{ac}+p_{ag}+p_{at}+p_{cc}+p_{cg}+p_{ct}++p_{gg}+p_{gt}+p_{tt})} = (p_{ac} + p_{ag} + p_{at} + p_{cg} + p_{ct} + p_{gt})$. For humans this is typically around 10/10,000 or .001 (the default). If all pairs of nucleotides were equally likely this would have a value of 0.60 .

`-phee` represents the background GC content. Basically $\frac{(p_c+p_g)}{(p_a+p_c+p_g+p_t)} = (p_c + p_g) = \phi$. Typical values for humans are around 0.4, the non-biased default is 0.5 .

`-ka` Transition / transversion ratio. This is $\kappa = \frac{(p_{ag}+p_{ct})}{(p_{ac}+p_{gt})}$. A typical value for humans is 5.0, but this can vary widely. The non-biased default is 1.0 .

As the primary output (.SNP) is verbose, it is often useful to filter out those positions that are highly likely to match the reference genome. The following flags help determine which positions get output to the .SNP file and which positions are ignored (or optionally output to the .nSNP digest).

2.4 Output Filtering

```
-mq : Min MAQ SNP              : 1 [range 0-255]
```

Output any position that MAQ would call with a phred score of >= -mq. As this is a phred score, a value of 1 means output any position that the MAQ SNP caller thinks has a greater than $1 - 10^{-\left(\frac{1}{10}\right)} \approx 20.6\%$ chance of being a SNP. Setting this to 0 disables use of MAQ calls as a criteria for SNP filtering, and BSNP relies only on the `-mp` parameter (below)

```
-mp : Min B SNP Probability    : 0.001000 [range 0.0-1.0]
```

Output any position where the posterior probability of failing to be a homozygous match for the reference sequence is $\geq$ this probability. Setting this to 0.0 causes all positions to be printed in verbose format to the SNP file. To only view high confidence SNP's, set this to a value near 1.0 (say .999) , to capture anything that might not match the reference, set this to a value near 0.0 (say 0.001). Essentially setting this to a small positive value insures that only positions that are high confidence matches to the reference, according to BSNP are relegated to the NoSNP file. This does not affect probability calculations, only which positions are output to SNP versus NoSNP output streams.

2.5 Correlated observation correction

```
-th : Theta (0-1.0,1.0=off)    : 0.850000   [range 0.0 to 1.0]
-tm : ThetaMin (0-1.0,0=off)   : 0.000000   [range 0.0 to 1.0 ]
-ti : Invert Theta Sorting     : 0          [ 0 or 1]
```

To account for possible correlation in observed data (perhaps due to unbalanced PCR duplicates), this data down-weights multiple sightings of large numbers of nucleotides at the same position. This is done by sorting the observed nucleotides at each position, in order of *decreasing* sequencing quality, and then exponentiating the probability of sequencing error by $\theta(i) = \theta^{i-1}$ where $1 \leq i \leq \#\ of\ reads\ at\ this\ position$. Note that $0 \leq \theta \leq 1.0$ so this exponentiation raises the probability of error towards 1.0, progressively decreasing the impact of successive nucleotide observations. The `ThetaMin` parameter allows us to set a floor for $\theta(i) = \max(\theta_{min}, \theta^{i-1})$ so nucleotides late in a deeply sequenced position are not entirely ignored. To disable this feature set `-th 1.0`. Setting the `-ti 1` allows BSNP to match the behavior of the SOAP caller which sorts nucleotides in order of *increasing* call quality, providing the lowest penalty for the lowest quality nucleotides sequenced. The value of 0.85 was empirically determined by the authors of MAQ, and was empirically verified on our 8 genome human data set, it does not have a theoretical basis.

2.6 OUTPUT FLAGS

The following flags enable or disable various output features. Each requires a 0 / 1 argument as well as the argument indicator.

```
-si : Dump Summary Info        : 1
-ns : Dump NonSNP Info         : 1
```

A value of 1 enables the output, a value of 0 disables it.

```
-ig : Ignore Invalid Bases     : 1
```

Removed any read flagged as 'N', or with read quality 0 or alignment quality 0 from the input stream at read time. These bases should not have a numerical effect on the output, but do slow processing.

```
-v  : Verbose Operation        : 0
```

Occasionally print the number of Pileup lines processed to stderr. This may corrupt .nSNP file if that is also sent to stderr. Do not set the verbose mode if you are sending .nSNP digest to stderr. Also, `-v 1` is required for the output of some debugging information, below.

```
-d  : Dump Debugging Info      : 0
```

For debugging purposes only, a flag that allows a developer to take action, if they so choose. Currently periodically dumps summary statistics to stderr.

```
-pb : Print BSNP Call Char     : 1
```

For ease of processing, creates an additional output column in the .SNP file that displays the genotype with the highest posterior probability. It does not affect probability calculations. Setting this parameter to 0 inhibits printing of this character.

2.7 TECHNOLOGY SPECIFIC ERROR MODELING

```
-st : Sequencing Technology    : 0
     0 = UNKNOWN, 1 = Illumina\Solexa, 2 = Sanger,
     3 = 454, 4 = SOLiD (Nuc Space)
```

BSNP is capable of taking into account known biases in sequencing technology to generate more accurate results. It is known that, in the event of a sequencing error, not all alternative bases have equal probability. By observing the distribution of nucleotides aligned to high confidence homozygous calls, but inconsistent with the consensus call we generate an technology-specific expected distribution for sequencing errors, that is $P(CorrectBase = X \mid CalledBase = Y \text{ ^ } CalledBase \neq CorrectBase)$. This can be incorporated into the posterior probability distribution via the sequencing quality value, which predicts the probability that a nucleotide in a short read is has

been incorrectly called. If this option is set to 0, the probability of each non-called base, in the event of an error, is set to a uniform $\frac{1}{3}$. The values for Illumina were based on the aggregate of 5 complete human genomes, those for Sanger, 454 and SOLiD were each based on a different complete human genome. This is an experimental method that has not been validated and should be used for comparison

3 OUTPUT FILES:

Three output files can be generated:

.SNP – Primary output, contains the chromosome name, position, reference call, MAQ call, MAQ quality values, BSNP posterior genotype probability distribution and data likelihood so complete joint of data and agenotype can be reconstructed. Also, individual nucleotide characters and qualities are provided. Optionally the highest posterior genotype is provided in a spate column. Fields are tab separated.

.nSNP – Digest of reference matches. Most called positions match the reference with high confidence. If this digest is selected, then a line is provded in the digest for each contiguous sequence of observations that matches the reference with proscribed quality (see –mq and –mp options). Each line provides the chromosome name, starting position and length of the sequence, and 2 subsequent strings each with one character per covered position. The first provides phred encoded probabilities of NOT matching the reference, the second a capped count of the number of read depth (optional).

.SUM – summary file containing aggregate information about number of lines processes, count of nucleotides observed, and NLL scores for observed data (optional).

4 EXAMPLES OF OUTPUT FILE LINES:

4.1 .SNP FILE – POSSIBLE POLYMORPHISM GENOTYPE DISTRIBUTION.

Fields: 1-9

#ChrName	ChrPos	RefCall	MAQcall	#BSNPcall	MAQcallQ	MAQsnpQ	MAQrmsQ	P(Data)_NLL10
chr1	3165	C	Y	C	6	6	21	9.1886e+00

Fields: 10-19

P(G=AA)	P(G=AC)	P(G=AG)	P(G=AT)	P(G=CC)	P(G=CG)	P(G=CT)	P(G=GG)	P(G=GT)	P(G=TT)
1.0994e-12	2.8125e-08	2.8680e-17	7.2353e-15	8.0574e-01	3.0653e-08	1.9426e-01	5.5184e-16	3.8753e-16	5.1014e-10

Fields: 20-29

P(D|G=AA)	P(D|G=AC)	P(D|G=AG)	P(D|G=AT)	P(D|G=CC)	P(D|G=CG)	P(D|G=CT)	P(D|G=GG)	P(D|G=GT)	P(D|G=TT)
2.0545e+01	1.2961e+01	2.1953e+01	1.9551e+01	8.6799e+00	1.2924e+01	6.1221e+00	2.3844e+01	2.0822e+01	1.7878+01

Fields: 30-34

NumBadReads	NumGoodReads	Reads	ReadQual_Q	AlignQual_Q
0	21	TCCTCTCTTTCCCCCCCCCCC	:24;493<<90<4<,<;;;3+	:)//7)88(80888:8:8888

1	ChrName	Chromosone name
2	ChrPos	Position of Chromosome, 0 based.
3	RefCall	Reference Call, not used in calculations, but inherited from Pileup file
4	MAQcall	Provided from Pileup File
5	BSNPCall	Optional, genotype with highest posterior probability
6	MAQcallQ	From Pileup, phred based quality of specific call made by MAQ
7	MAQsnpQ	From :Pileup phred based likelihood of SOME SNP at this position (not necessarily the MAQ call)
8	MAQrmsQ	from Pileup RMS sequencing quality of nucleotides at this position
9	P(Data)_NLL10	The negative log likelihood (base 10) of the sequenced data likelihood according to the BSNP model, derived as $\sum_{g \in Genotypes}[P(Data\|g)P(g)]$.
10-19	P(G|Data)	Posterior probability of genotypes at this position, given the sequence data.
20-29	P(Data|G)	Data Likelihood at this position given the genotype (negative log likelihood, base 10)
30	Num Bad Reads	Number of reads that had 0 sequence quality, 0 alignment quality or a non ACGT symbol.
31	Num Good Reads	Number of reads that had positive alignment & sequence quality and were a canonical symbol (ACGT)
33	Reads	Nucleotides data at this position
34	ReadQuality	Phred encoded nucleotide sequence quality
35	Align Quality	Phred encoded quality of short read from which this nucleotide was drawn.

4.2 .nSNP FILE – NONSNP DIGEST.

```
#ChrName    ChrPos      NumContig   NumValidReads                ProbSNP-5Lg10(p)
chr1        2985        28          (('()******))('&&&&&&&&&%%%%  9989;<5<8<<:;986666666665555
```

1	ChrName	Chromosone name
2	ChrPos	Position of Chromosome, 0 based.
3	NumContig	Number of positions covered in this line.
4	NumValid	Each value ascii character -33 is the number of reads covering this position, capped at 93.
5	ProbSNP	Phred-style encoded probability of a SNP at this location, to recover probability take ASCII value of character v and apply $P(is\ a\ SNP) = 10^{-\left(\frac{v-33}{5}\right)}$. I divide by 5 rather than the traditional 10 to cover a larger range (down to 2.5E-19), at a cost in precision (largest non-1.0 value is .63). As this is a digest of positions that were not included in the SNP file, we already know that this is not likely to be a SNP, we just want to know how deeply unlikely it is.

4.3 .SUM FILE – GENOME SUMMARY DATA

```
Pileup - Lines Read – Tot:    Read: Total lines in Pileup File, should match linux wc command.
Pileup - Lines Skipped:       Read: Indel lines, header etc…
Pileup - Lines Processed  :   Read: Should be Tot – Skipped

Caller - Lines Processed      Caller: Should = Pileup - Lines Processed
Caller - Lines Voided         Number of lines with no data, or where all data has 0 quality.
Caller - Lines Assessed       Processed – Voided
Caller - SNPs  called         Genomic positions written to SNP file
Caller - nSNPs called         Genomic positions written to .nSNP file (if requested by user)
Caller - nSNPs from Void      A line may be written to the .nSNP file for two reasons. Either
                                    1) The data could rejected SNP hypothesis at a given confidence
                                    2) There was no valid data
                                    This can be confusing because, based solely on the priors, P(Geno | No Data) could
                                          fail to reject the SNP hypothesis due to low genotype confidence,
                                          and thus would be included in the SNP file, filling that file
                                          with a large volume of useless data. This line counts how many
                                          lines were included in .nSNP because they met the "no valid data" condition.

Nucs - Total -                Number of each nucleotide character in input file.
Nucs - Unvoided               Number of VALID nucleotide characters in input file (sequencing quality >0)
Nucs - Partial                Number of valid reads prorated by alignment quality. (1-P_ErrAlign)
Nucs - Effective              Number of valid reads prorated by alignment & Seq quality. (1-P_ErrAlign)(1-P_ErrSeq)

Geno - Counts                 Sum of probabilities of each genotype, over all "Caller Lines Assessed"
Data NLL10 – Total            Data Negative Log Likelihood (base 10) over all "Caller Lines Assessed"
Data NLL10 – PerNuc           NLL Total / nucs_partial.
```

5 History

BSNP originated as a graduate-level class project for Adam Siepel's[xvii] BTRY6840 class in 2009[xviii]. It was utilized to combine human genomic data from a variety of sequencing technologies under a uniform sequencing and calling pipeline for human demographic inference [[xix]]. Researchers using a variety of aligners and conventional SNP callers tended to use differing heuristics to achieve a low false positive rate among SNP calls, but the differences in methodologies obscured subtle signals needed to infer demographic properties. BSNP, applied to a uniform realignment pipeline based on the BWA aligner was used to mitigate these effects. Version 2.05a (2010-May) was used in the reference paper [see References: "Bayesian inference of ancient human demography from individual genome sequences"], while the first publically released version was 2.14 (2012-May), which incorporated several bug fixes as well as sequencing technology specific bias compensation. Please contact the author with requests for updates or bug reports. As of 2012 code is lightly, but consistently, maintained but not under active development.

Author: Brad Gulko, Department of Computer Science, Cornell University, Ithaca, NY, USA. bgulko@cs.cornell.edu, bg279@cornell.edu, bgulko@LeptonCorp.com. http://www.cs.cornell.edu/~bgulko/, http://compgen.bscb.cornell.edu/~bgulko/, 2012.

6 Appendix

Diploid Genotypes – the 16 combinations of the four canonical nucleotides {A, C, G, T}

Un-phased Diploid Genotypes – the 10 unordered combinations of the four canonical nucleotides {AA, AC, AG, AT, CC, CG, CT, GG, GT, TT}

7 Software References

BSNP

Publication: http://www.nature.com/ng/journal/v43/n10/abs/ng.937.html.
In particular, the supplementary material at http://www.nature.com/ng/journal/v43/n10/extref/ng.937-S1.pdf section S1.3 "BSNP Method for Genotype Inference" and S1.3.3 for the correlated data compensation.
Ilan Gronau, Melissa J Hubisz, Brad Gulko, Charles G Danko & Adam Siepel, "Bayesian inference of ancient human demography from individual genome sequences", Nature Genetics Volume: 43, Pages: 1031–1034 (2011) doi:10.1038/ng.937.

SAMtools

Code: http://samtools.sourceforge.net/
Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

MAQ

Code: http://maq.sourceforge.net/
Publication: http://genome.cshlp.org/content/18/11/1851
Publication: http://genome.cshlp.org/content/18/11/1851.full.pdf

Heng Li, Jue Ruan, and Richard Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". Genome Res. 2008. 18: 1851-1858.

BWA

Code: http://bio-bwa.sourceforge.net/bwa.shtml
Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60. [PMID: 19451168]

SOAP2 / SOAPsnp

Code : http://soap.genomics.org.cn/soapaligner.html
Publication : http://bioinformatics.oxfordjournals.org/content/25/15/1966.abstract
Ruiqiang Li, Yingrui Li, Xiaodong Fang, et al. (2009) "SNP detection for massively parallel whole-genome resequencing" (2009) Genome Res. , doi:10.1101/gr.088013.108

8 ADDITIONAL REFERENCES

[i] For both the SAM and BAM file formats see the Samtools reference, below. As of 2012-May, both are incorporated into the SAM specification at http://samtools.sourceforge.net/SAM1.pdf.

[ii] Samtools: Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]. http://samtools.sourceforge.net/

[iii] SAMtools pileup command: http://samtools.sourceforge.net/cns0.shtml. The command form `samtools pileup -csf hg18.fasta reads.bam > pileup.txt` was used to generate the output format required by BSNP. On 2010 era hardware, this took around 6 hours to complete.

[iv] Ruiqiang Li, Yingrui Li, Karsten Kristiansen and Jun Wang, "SOAP: short oligonucleotide alignment program", Bioinformatics (2008) 24 (5): 713-714. doi: 10.1093/bioinformatics/btn025. URL at http://bioinformatics.oxfordjournals.org/content/24/5/713.short . Software at http://soap.genomics.org.cn .

[v] International HapMap project, http://www.hapmap.org/ or http://hapmap.ncbi.nlm.nih.gov/ . See also: The International HapMap Consortium. "The International HapMap Project." Nature, 426, 789-796. 2003. And also, Thorisson, G.A., Smith, A.V., Krishnan, L., and Stein, L.D. "The International HapMap Project Web site." Genome Research,15:1591-1593. 2005.

[vi] Adam Siepel (2012) Associate Professor, Biological Statistics & Computational Biology, Director of Graduate Studies Computational Biology, Associate Director Cornell Center for Comparative and Population Genomics, Cornell University. 102E Weill Hall, Cornell University, Ithaca, NY 14853. Phone: 607-254-1157 http://compgen.bscb.cornell.edu/~acs

[vii] Bzip2. Open source data compressor available on most standard Linux distributions, or from http://www.bzip.org/index.html more information available from Wikipedia at http://en.wikipedia.org/wiki/Bzip2 .

[viii] WinRAR Free trial version available at http://en.wikipedia.org/wiki/WinRAR more information at http://en.wikipedia.org/wiki/WinRAR .

[ix] 7Zip - Free open source software available at http://www.7-zip.org/ more information available on Wikiepdia at http://en.wikipedia.org/wiki/7zip.

[x] Heng Li, Jue Ruan, and Richard Durbin, "Mapping short DNA sequencing reads and calling variants using mapping quality scores", Genome Reserarch, 2008. 18: 1851-1858. http://genome.cshlp.org/content/18/11/1851.full . Software at http://maq.sourceforge.net/index.shtml .

[xi] Ruiqiang Li, Yingrui Li, Karsten Kristiansen and Jun Wang, "SOAP: short oligonucleotide alignment program", Bioinformatics (2008) 24 (5): 713-714. doi: 10.1093/bioinformatics/btn025. URL at http://bioinformatics.oxfordjournals.org/content/24/5/713.short . Software at http://soap.genomics.org.cn .

[xii] This can be downloaded as a single file at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/chromFa.zip, or via individual chromosomes at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/chromosomes/ . To generate a single hg18.fa file from individual chromosomes, decompress each chromosome's .fa file. The results are simple text files that may be concatenated together to generate a single hg18.fa (ie. cat chr*.fa > /tmp/hg18.fa). The resultant file is about 3.5 GB uncompressed.

[xiii] As provided by http://samtools.sourceforge.net/pileup.shtml "Pileup format is first used by Tony Cox and Zemin Ning at the Sanger Institute. It describes the base-pair information at each chromosomal position. This format facilitates SNP/indel calling and brief alignment viewing by eyes."

[xiv] See "Consensus Calling" in http://samtools.sourceforge.net/cns0.shtml .

[xv] For definition of ASCII see http://en.wikipedia.org/wiki/ASCII . ASCII is a simple encoding of a subset of printable characters into 1 byte values where visible characters span the numerical range 33 – 126, inclusive.

[xvi] See: http://en.wikipedia.org/wiki/Phred_quality_score or http://www.phrap.org/phredphrapconsed.html

[xvii] Adam Siepel, Cornell University, Associate Professor, Biological Statistics & Computational Biology, Director of Graduate Studies, Computational Biology http://compgen.bscb.cornell.edu/~acs/

[xviii] Cornell University, "BTRY 4840/6840: Computational Genomics", 2009, URL= http://compgen.bscb.cornell.edu/btry4840/index.php/Main_Page

[xix] Ilan Gronau, Melissa Hubisz, Brad Gulko, Charles G. Danko, and Adam Siepel, "Bayesian inference of ancient human demography from individual genome sequences", Nat Genet. 2011 September 18; 43(10): 1031–1034. doi: 10.1038/ng.937. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245873/ . Particularly see supplemental information "S1 Genotyping Pipeline" regarding BSNP.