Table 1. Data Sources and Software Tools used for Generating SNP2TFBS

Source Data		
SNP catalog (1000 Genomes)	VCF file from 1000 Genomes (version v5a.20130502), filtered for minor allele frequency > 0.01	[a]
Human reference genome (UCSC)	Human reference genome GRCh37/hg19 in FASTA format download	[b]
PWM collection (JASPAR/MEME)	JASPAR Core Vertebrate 2014 PWM from MEME motif database version 12.1	[c]
Gene annotation (RefSEQ, Annovar)	From RefSeq (version Feb 2016) as provided by ANNOVAR version v2016-02-0	[d]
External and in-house software tools		
vcf2diploid (v0.2.6)	Mapping alleles from vcf in reference to generate alternate genome	[e]
GATK (v3.6)	Liftover variants from reference to alternate assembly	[f]
Samtools (v0.1.14)	Generating genome assembly index file (required for Picard)	[g]
Picard (v1.131)	Building genome dictionary for GATK	[h]
ANNOVAR (v2016-02-01)	Annotating variants of interest with refGene annotation (version Feb 2016)	[d]
PWMTools (v1.0.0)	Scanning a whole genome with a PWM (**matrix_scan**), and determining the P-value threshold for PWM scores (**matrix_prob.pl)**	[i]
Scripts available from the SNP2TFBS FTP site		
makeAltGenome.sh	Wrapper to generate alternate genome assembly	
makeSNP2TFBS.sh	Wrapper for scanning genomes with PWMs and output both mapped files (single PWM format) and SNP2TFBS master file	
makeDerivedFormats.sh	Wrapper to generate derived formats (bed, sga, and annotated) from both SNP2TFBS single PWM and master files	
vcf_filter.pl	Input is a single or multi-sample vcf file (with SNP and indels) and output is a single column vcf file with AF≥0.01 and alleles with highest AF	
snp_table.pl	Merging and indexing variants for both the genome assemblies	
variantPWMmatch.pl	Mapping variants in PWM sites of both the genomes	
filterVariantPWMmatch.pl	Merging and filtering PWM sites with variants from both the genome assemblies	
mergeMappedFiles.pl	Merging mapped files for all the factors and output a single file with PWM in rows sorted with absolute score difference between both genome assemblies	

a) 1000 Genomes: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/
b) hg19: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/
c) MEME motif database: http://jaspar2014.genereg.net/html/DOWNLOAD/
d) ANNOVAR: http://annovar.openbioinformatics.org/en/latest/ [5]
e) vcf2diploid: (http://alleleseq.gersteinlab.org/vcf2diploid_v0.2.6.zip) [1]
f) samtools: https://sourceforge.net/projects/samtools/files/samtools/ [2]
g) picard: https://github.com/broadinstitute/picard/releases/download/1.131/picard-tools-1.131.zip [3]
h) GATK: https://software.broadinstitute.org/gatk/download/ [4]
i) PWMTools: https://sourceforge.net/projects/pwmscan/

References:

[1] Rozowsky J. et al. AlleleSeq: analysis of allele-specific expression and bin ding in a network framework. Mol Syst Biol. 2011

[2] The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA, 2010 GENOME RESEARCH 20:1297-303

[3] 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

[4] http://picard.sourceforge.net

[5] Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research, 38:e164, 2010