File extension	Description	Links	Programs commonly used in/produced by
.fa	Fasta files are files that contain a nucleotide or protein sequence, generally without quality scores. This will typically be a high confidence nucleotide sequence or reference assembly. The name of each sequence included is preceded by the > symbol, followed by the sequence.	https://zhanglab.ccmb.med.umich.edu/FASTA/	Many!
.fq/.fastq (also .fq.gz and .fastq.gz)	Read data from next-generation sequencing. There are four lines per read, including information such as the read name, the nucleotide sequences read and relevant quality scores	https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html	Mapping programs, de novo assembly programs, among other
.sam/.bam	Format for reads mapped to a reference genome including important information such as the read, mapping location, and mapping quality. The bam file is binary while the sam file is uncompressed.	https://genome.sph.umich.edu/wiki/SAM	bwa, STAR, stampy, and others
.vcf/.bcf	Variant file containing information on variant base pairs, mapping quality, depth, and a large number of base quality metrics relative to the reference genome used for mapping. The vcf file is uncompressed which the bcf file is the binary version. **Note: the specifics of the vcf format will vary from program to program**	https://www.ebi.ac.uk/training/online/course/human-genetic-variation-introduction/exercise-title/want-know-how-we-did-it	samtools/bcftools, GATK, and others

.g.vcf	Similar to a vcf but also containing information for invariant sites. **Note: the specifics of the vcf format will vary from program to program**		samtools/bcftools, GATK, and others
ped map	Compact file format containing pedigree (optional), phenotype (optional), and genotype data that is used by the plink program and several other programs. The ped file must be paired with a map file which gives the locations of SNPs	http://www.gwaspi.org/?page_id=145	plink
bed	bed files are useful file formats for storing interval information for a particular genome assembly. The first column contains the linkage group/scaffold/chromosome name, the second column contains the feature start position, the third column contains the feature end position. Additional columns after these first three are allowed by most programs	https://www.genomatix.de/online_help/help_regionminer/bedformat_help.html	Can be included to incorporate known features (e.g. repetitive regions, exons) in many programs
gff/gtf	Annotation files that contain information on specific features in the genome. A gtf is a special case of a gff file that contains information on gene structure (i.e. exon locations).	http://genomewiki.ucsc.edu/index.php/Genes_in_gtf_or_gff_format	Can be included as input with many programs, usually will be provided by a database (i.e. ensembl, ncbi, etc)
insnp	A file in the format: chromosome, position, basepair current, basepair to change to for use with		seqtk

	the seqtk program for updating fasta files		

Lab-specific file formats:

Ancestry-tsv format: posterior probabilities of ancestry for each individual and ancestry informative site in the genome

genotype format: hard-calls for ancestry in the same format as Ancestry-tsv files (2: homozygous parent 2 – *malinche*, 1: heterozygous ancestry, 0: homozygous parent 1 – *birchmanni*). We typically use a posterior probability cutoff of 0.9.

hybrid index file format: file including the estimated proportion of the genome from *malinche* and estimated ancestry heterozygosity