# Supplemental Note S1. Simulated Population Analysis

For evaluation we wrote a simulation script that generates a multi-sample VCF file with an arbitrary population structure. Briefly, the simulator simulates $F$ founder genotypes, that each contain on average $Normal(N, M)$ variants placed at random along the genome (the initial genome size is fixed at 100,000,000 bp). Then for each founder population, a collection of $Normal(S, T)$ individual samples are generated at random that contain the original founder variants plus an additional $Normal(X, Y)$ variants. Consequently, the expected total number of variants in the collection is $F * N + F * S * X$ variants. If $N > S$, then most of the variants will be shared within the population group, and if $S > N$, most variants will be unique to that sample. We emphasize this is not designed to simulate realistic pedigrees, but to examine the extremes of high or low levels of sharing among the individuals.

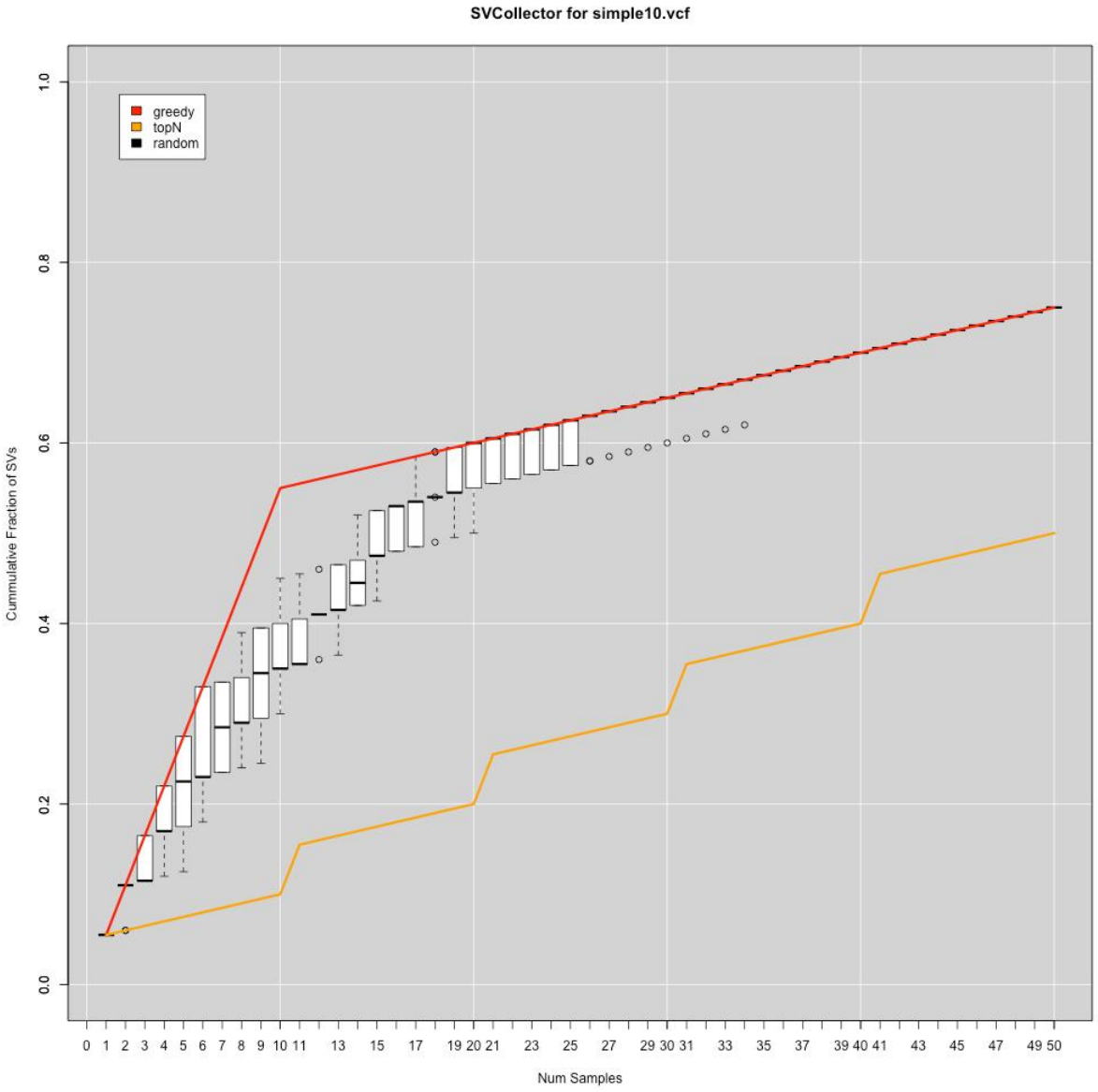
The code can be found at our GitHub page: https://github.com/fritzsedlazeck/SVCollector/tree/master/simul.

The first simulated population (simple10.vcf) has 10 founder genomes with exactly 1000 variants located at random. From each founder genome, 10 samples are simulated that contain the 1000 founder variants plus an additional 100 variants. The SVCollector curve for this sample is presented in **Supplemental Figure S1**. The sharp inflection point for the greedy curve at N=10 illustrates how the code realizes there are 10 founder populations. After these 10 populations have been sampled, the rate at which additional SVs are identified reduces to a much slower rate as these variants are only contained in individual samples.

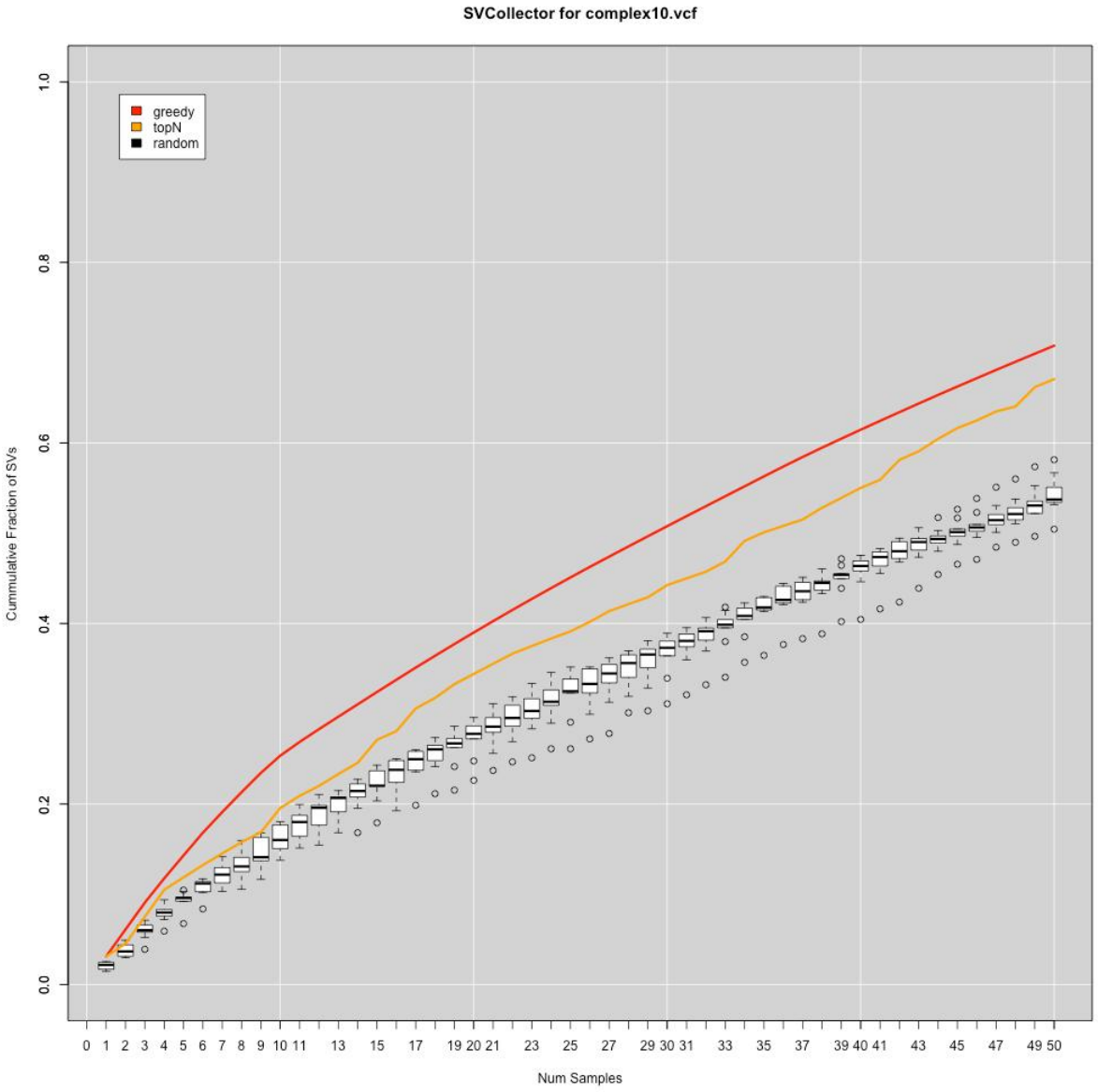
The simulator was executed like this: `./popsim.pl 10 10 0 1000 0 100 0 > simple10.vcf`

The second simulated population (complex10.vcf) also has 10 founder genomes that each randomly contain $Normal(500, 250)$ variants. From each founder population, $Normal(10, 5)$ samples are simulated that contain the founder variants, plus an additional $Normal(500, 250)$ variants unique to this sample. The SVCollector curve is presented in **Supplemental Figure S2**. Notice that despite having the same number of founder genomes, the curves are substantially different, and lack the inflection point at $N = 10$. This highlights how sample specific variants contribute at a similar level to the founder genotypes.

The simulator was executed like this: `./popsim.pl 10 10 .5 500 .5 500 .5 > complex10.vcf`




**Supplemental Figure S1:** SVCollector curve of a simple simulated population with most variants shared within the population group.

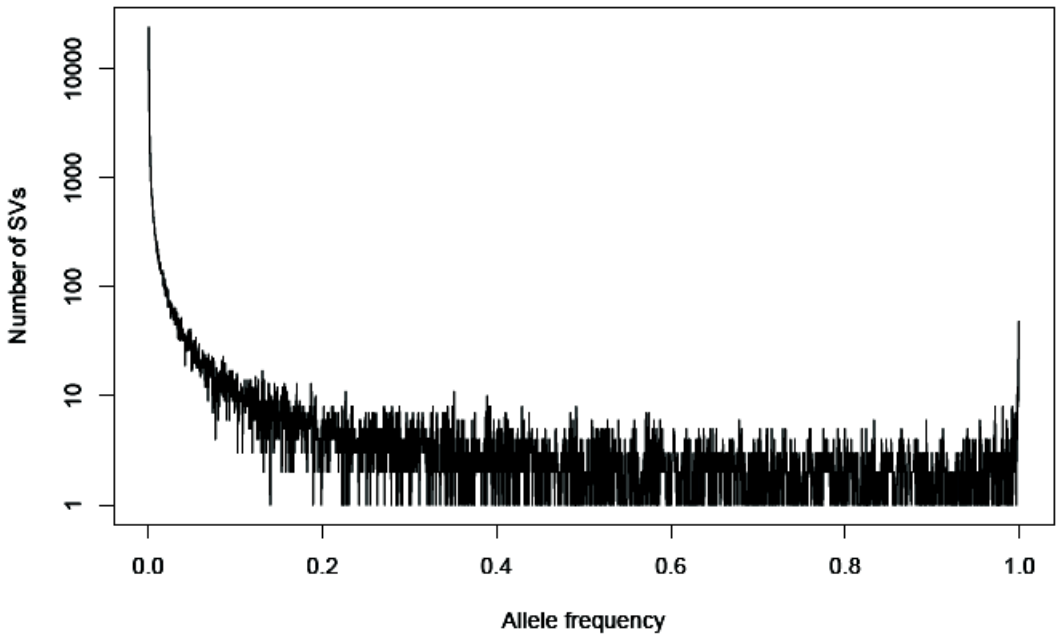



**Supplemental Figure S2:** SVCollector curve of a complex simulated population with most variants specific to the individuals.

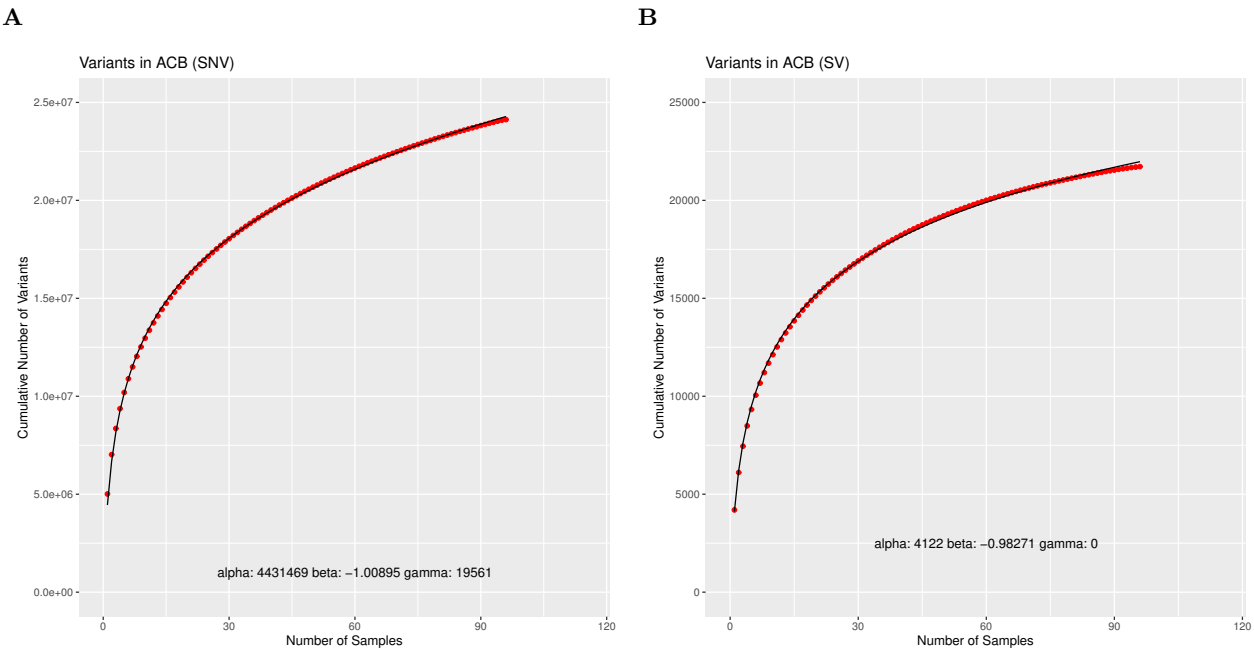
## Supplemental Note S2. 1000 Genomes Analysis

We downloaded the VCF file from the 1000 genomes FTP site: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz.

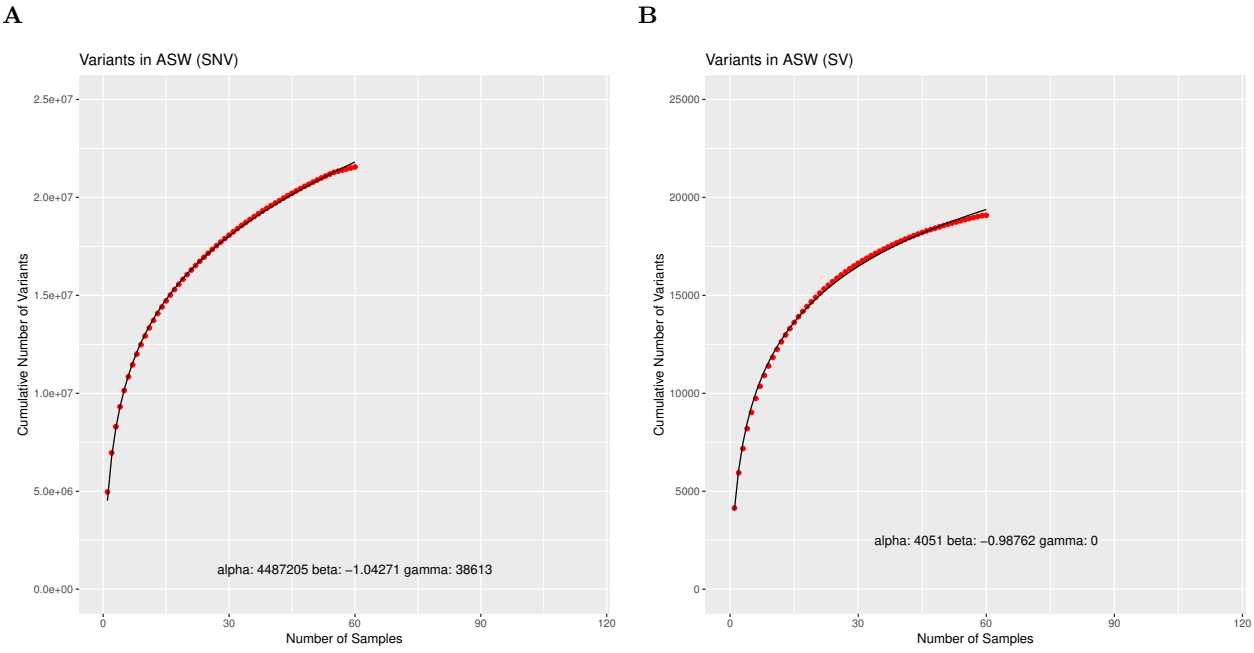
To further assess the robustness of our ranking using the *greedy* approach we sub-selected once 70% (46,589 SVs) of the SVs reported over the 1000 genomes and once 50% (33,278 SVs) of the SVs. Next, we compared the selection of these sub-selected lists of SVs to the original ranking based on all SVs. For the 70% subsampled set we observed that 7 out of the top 10 accessions are the same compared to the full text. This ratio remains similar for the top 100 ranked samples out of which 71 samples where the same. For the 50% subsampled VCF file we observed similar numbers where 7 out of 10 remained the same within the top 10 ranked samples. For the top 100 this ratio dropped to 59 out of the 100 samples. This illustrates the robustness of SVCollector when it comes to sensitivity issues for SV calling as the ranking remains very similar even with only 50% of the SVs present.



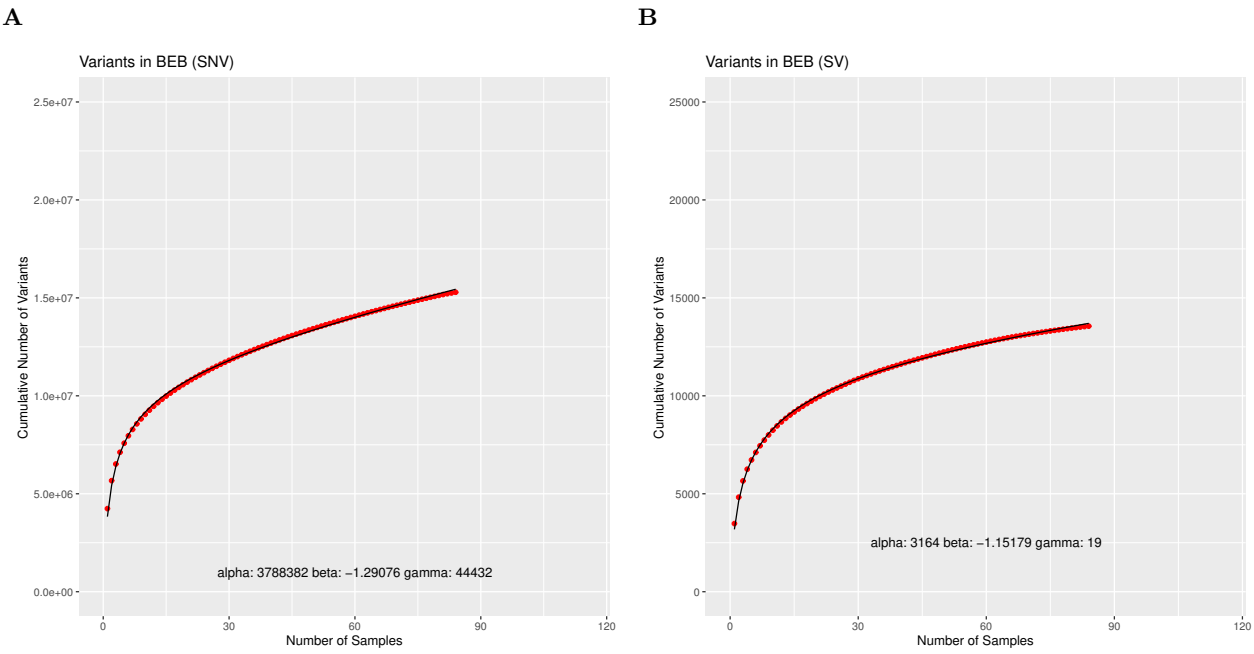

**Supplemental Figure S3:** Allele frequency distribution across the 1000 Genomes samples



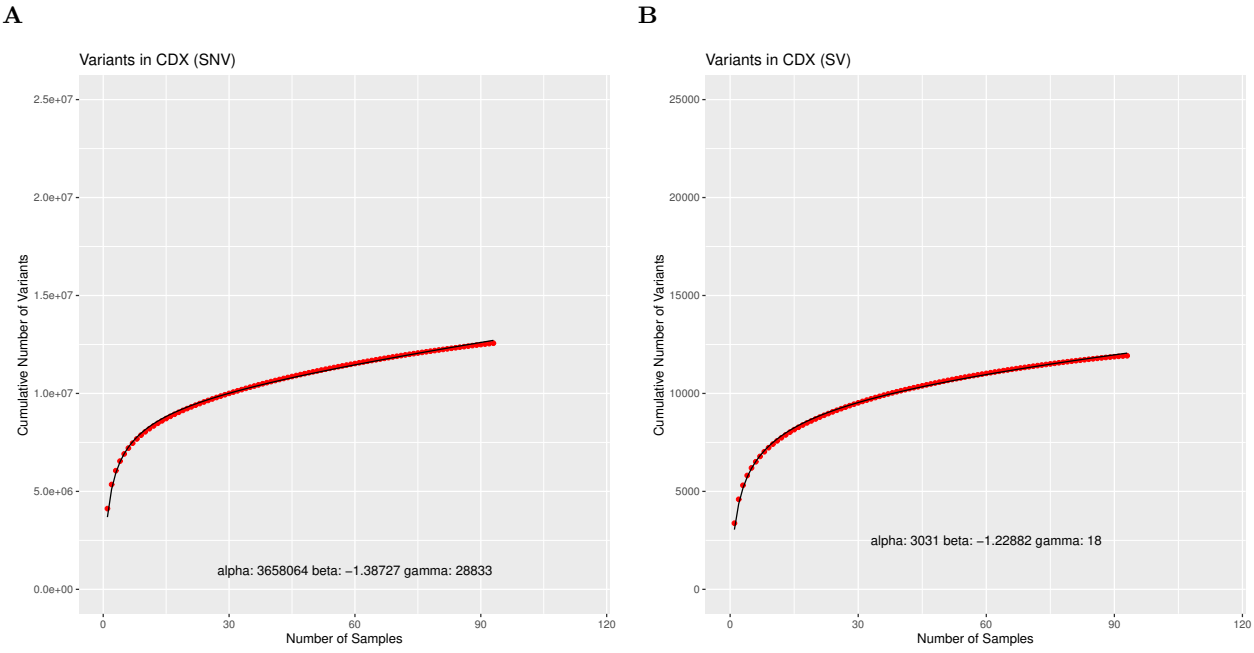

**Supplemental Figure S4:** SVCollector curves for the African Caribbeans in Barbados (ACB) population using **(A)** SNV data and **(B)** SV data.



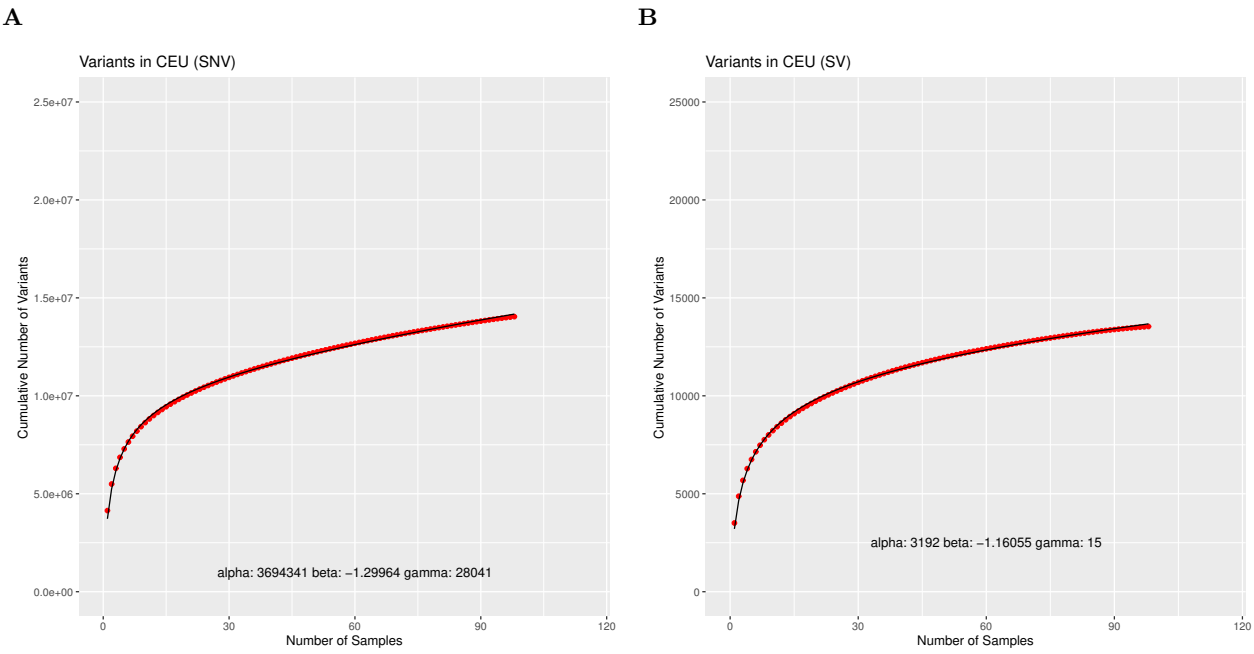

**Supplemental Figure S5:** SVCollector curves for the Americans of African Ancestry in SW USA (ASW) population using **(A)** SNV data and **(B)** SV data.



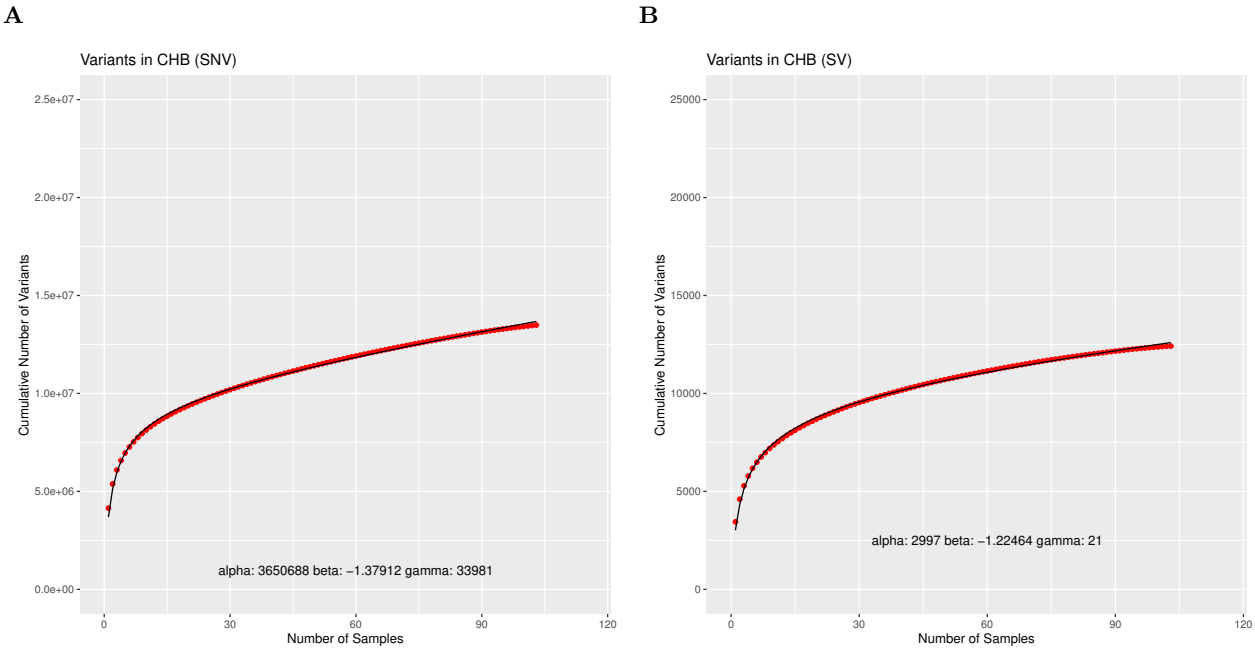

**Supplemental Figure S6:** SVCollector curves for the Bengali from Bangladesh (BEB) population using **(A)** SNV data and **(B)** SV data.



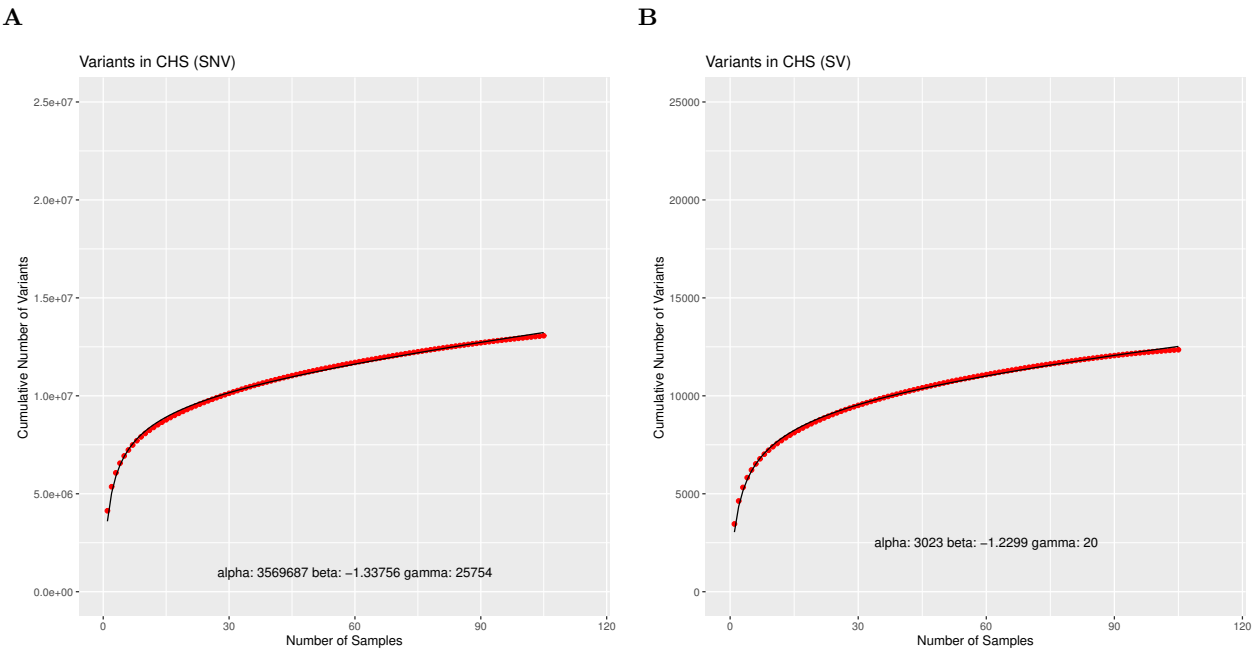

**Supplemental Figure S7:** SVCollector curves for the Chinese Dai in Xishuangbanna, China (CDX) population using **(A)** SNV data and **(B)** SV data.



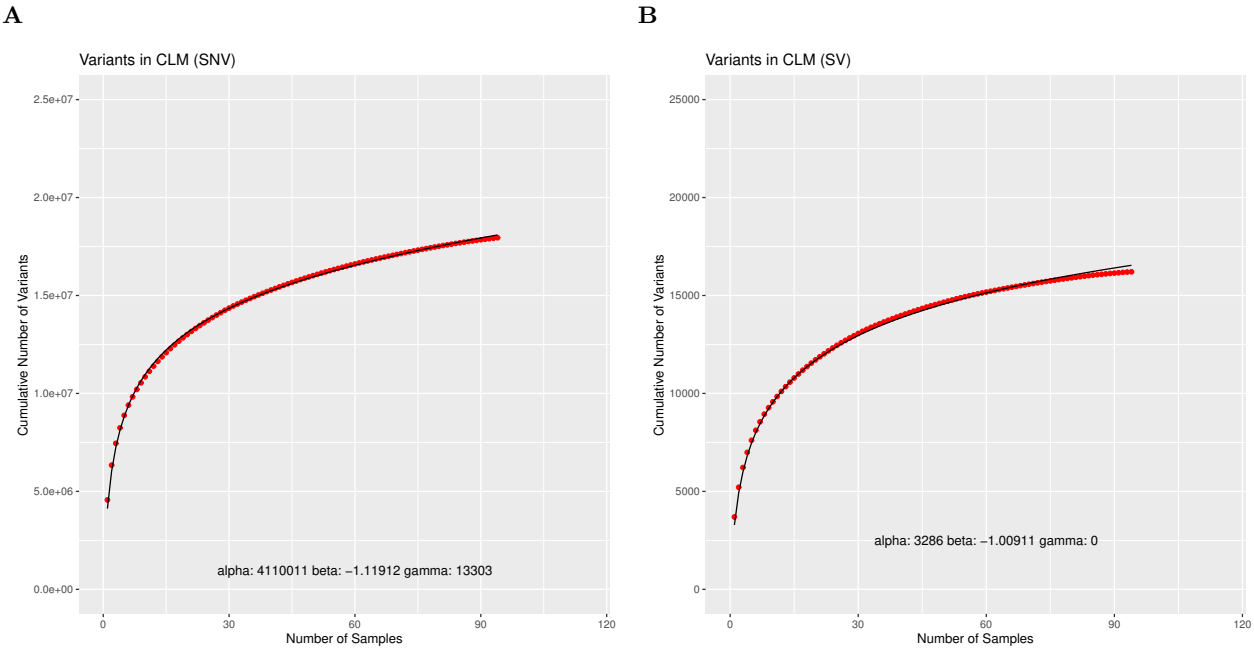

**Supplemental Figure S8:** SVCollector curves for the Utah Residents (CEPH) with Northern and Western European Ancestry (CEU) population using **(A)** SNV data and **(B)** SV data.



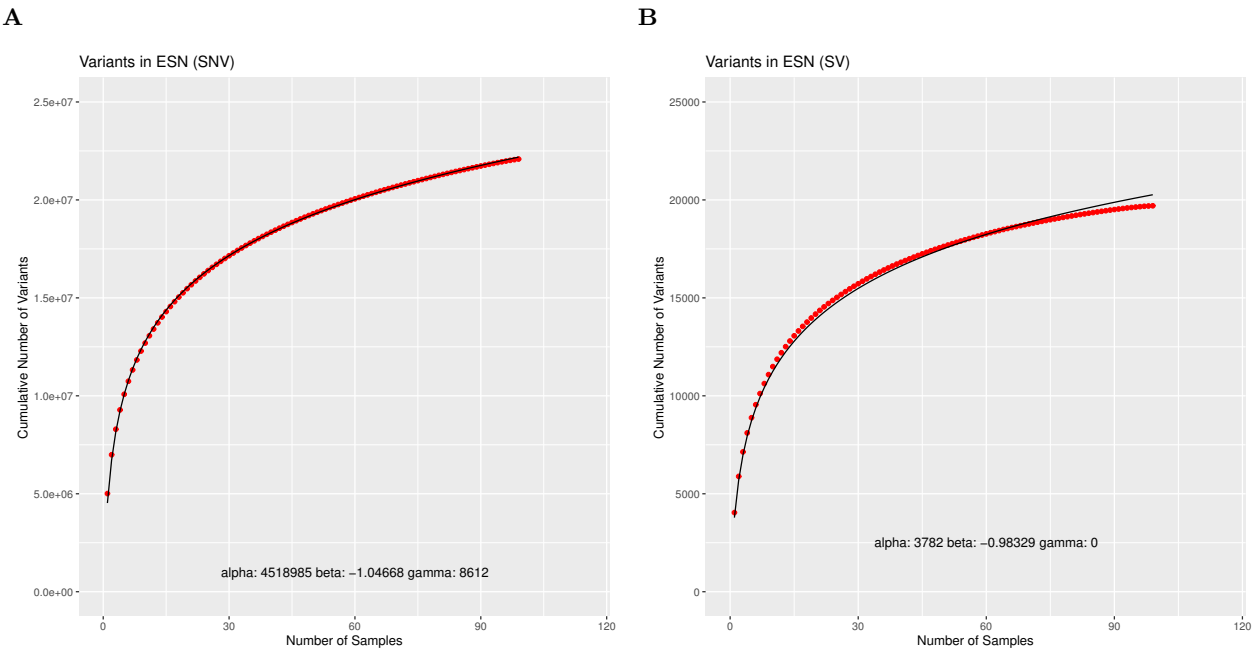

**Supplemental Figure S9:** SVCollector curves for the Han Chinese in Beijing, China (CHB) population using **(A)** SNV data and **(B)** SV data.



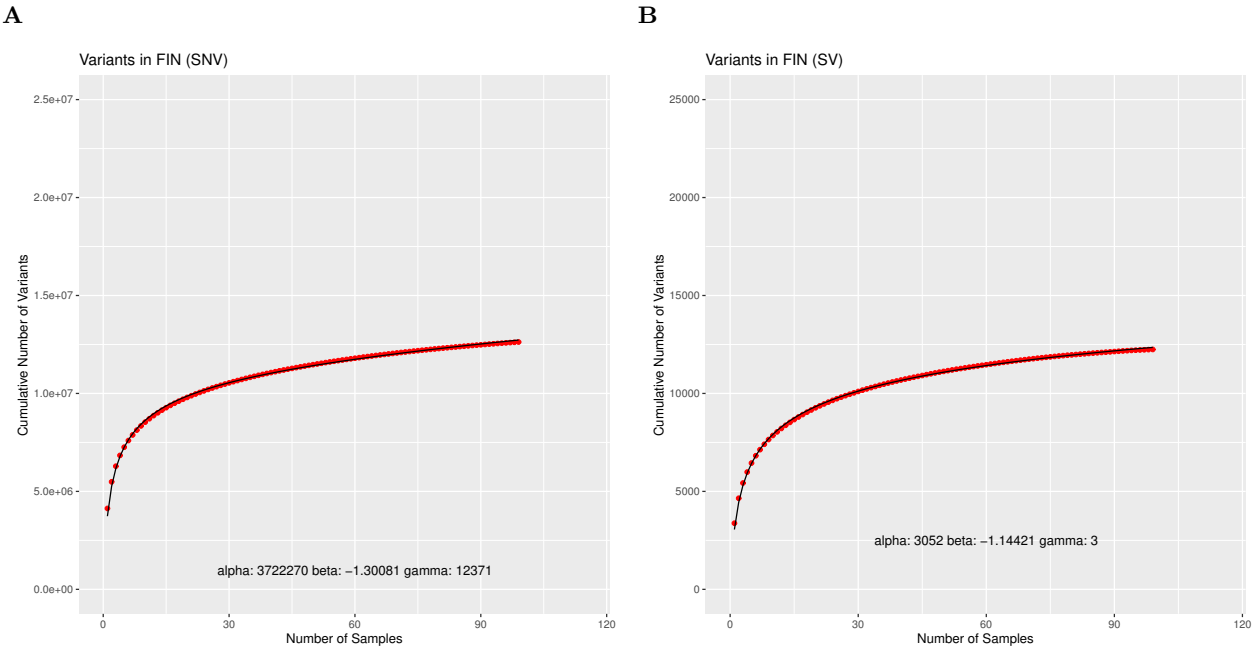

**Supplemental Figure S10:** SVCollector curves for the Southern Han Chinese (CHS) population using **(A)** SNV data and **(B)** SV data.



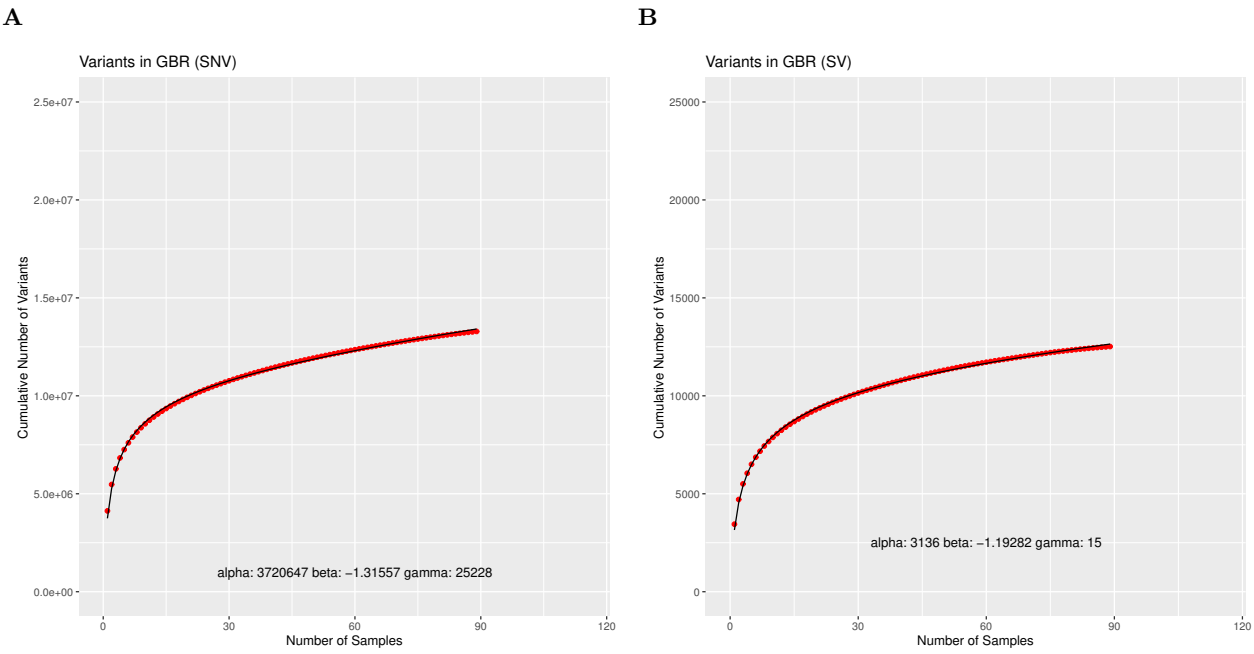

**Supplemental Figure S11:** SVCollector curves for the Colombians from Medellin, Colombia (CLM) population using **(A)** SNV data and **(B)** SV data.



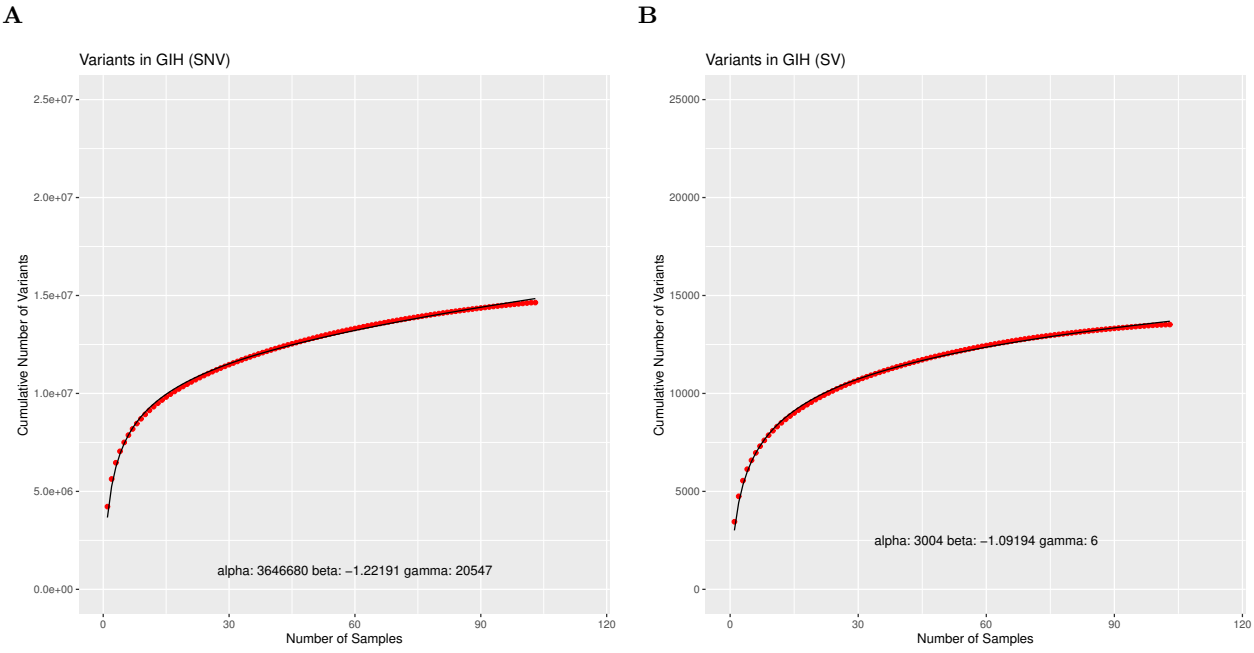

**Supplemental Figure S12:** SVCollector curves for the Esan in Nigeria (ESN) population using **(A)** SNV data and **(B)** SV data.



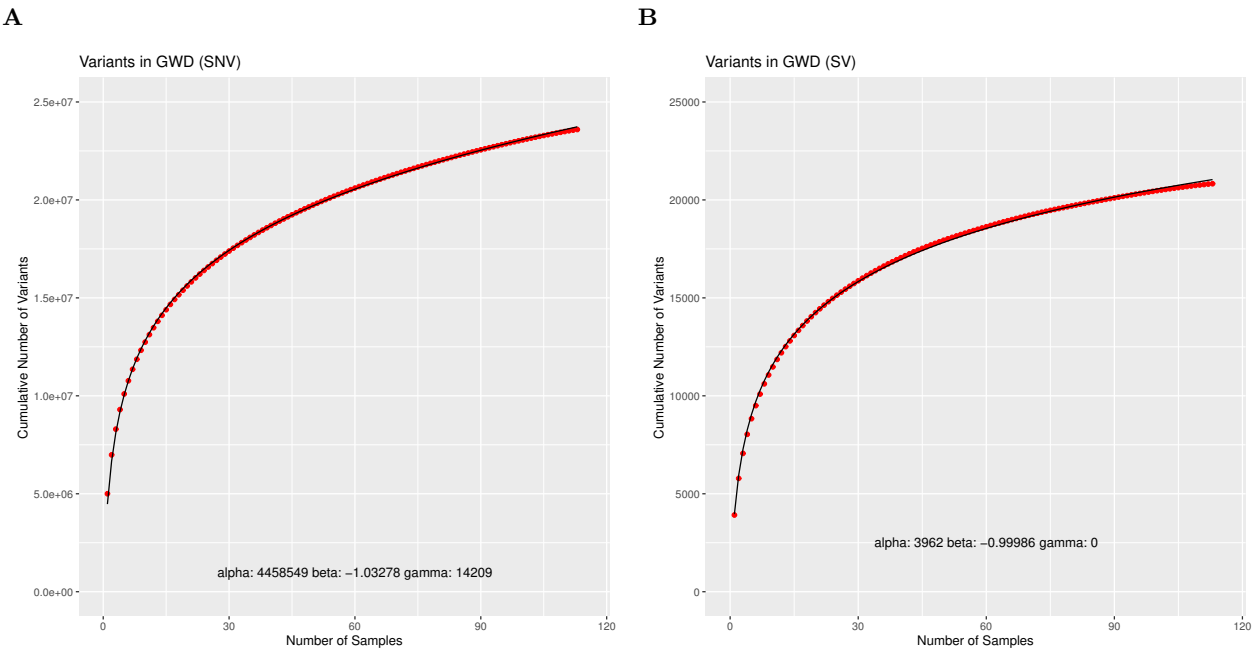

**Supplemental Figure S13:** SVCollector curves for the Finnish in Finland (FIN) population using **(A)** SNV data and **(B)** SV data.



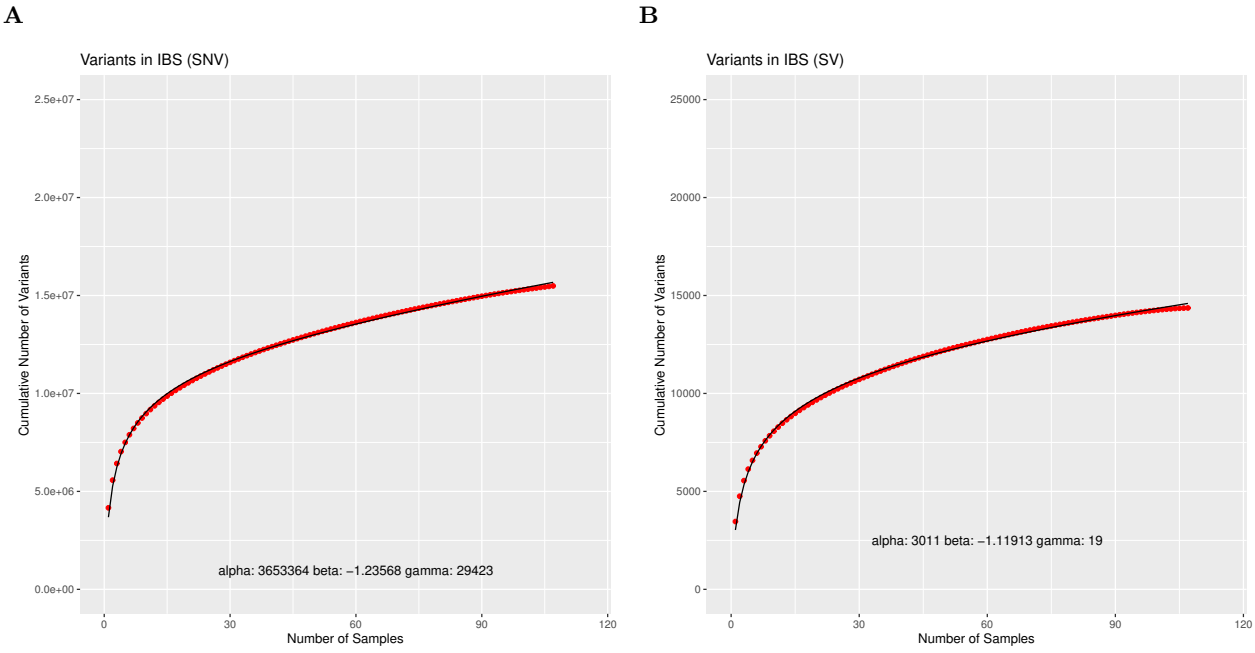

**Supplemental Figure S14:** SVCollector curves for the British in England and Scotland (GBR) population using **(A)** SNV data and **(B)** SV data.



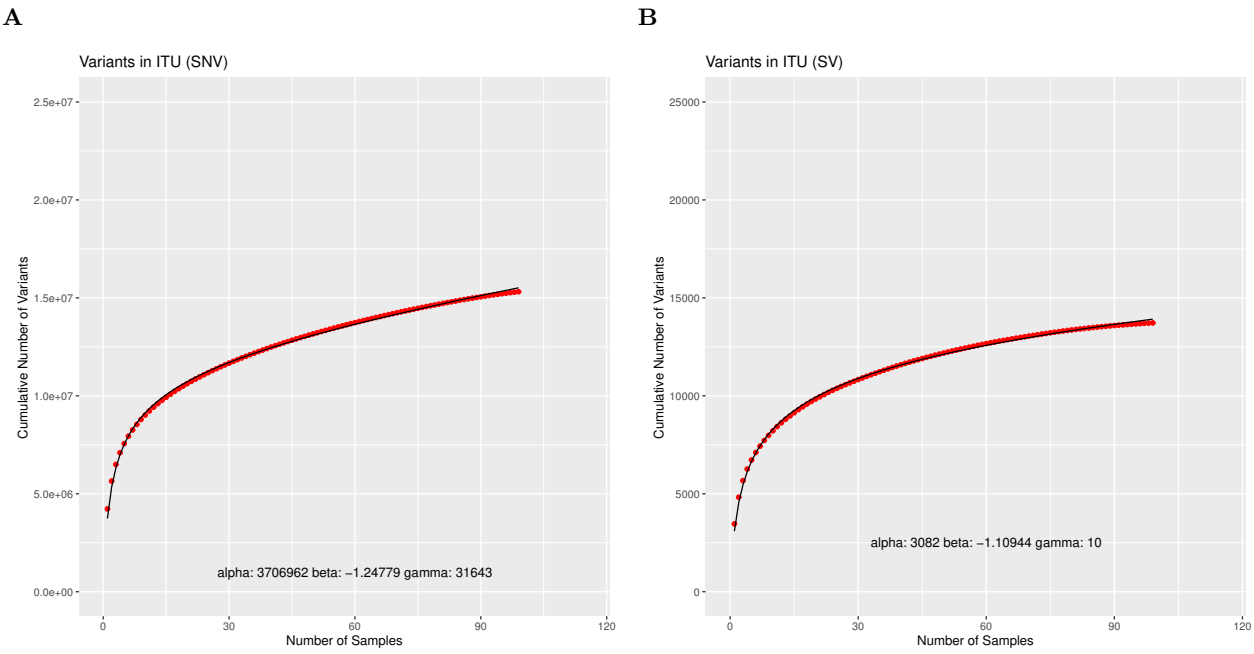

**Supplemental Figure S15:** SVCollector curves for the Gujarati Indian from Houston, Texas (GIH) population using **(A)** SNV data and **(B)** SV data.



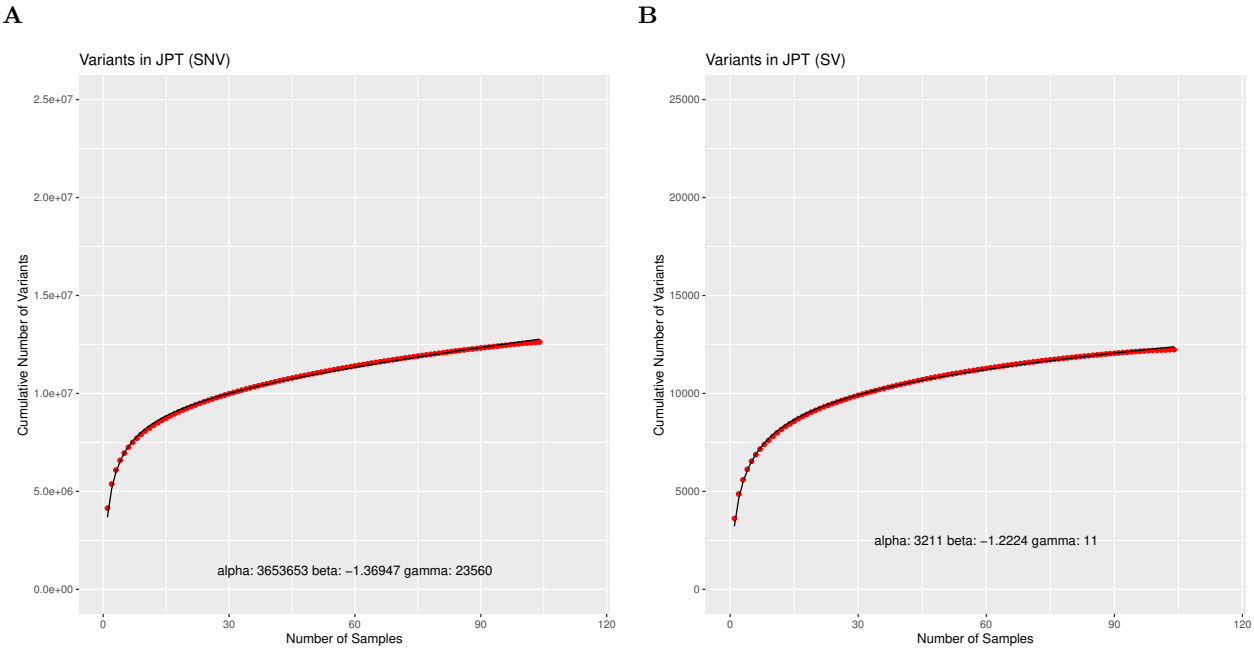

**Supplemental Figure S16:** SVCollector curves for the Gambian in Western Divisions in the Gambia (GWD) population using **(A)** SNV data and **(B)** SV data.



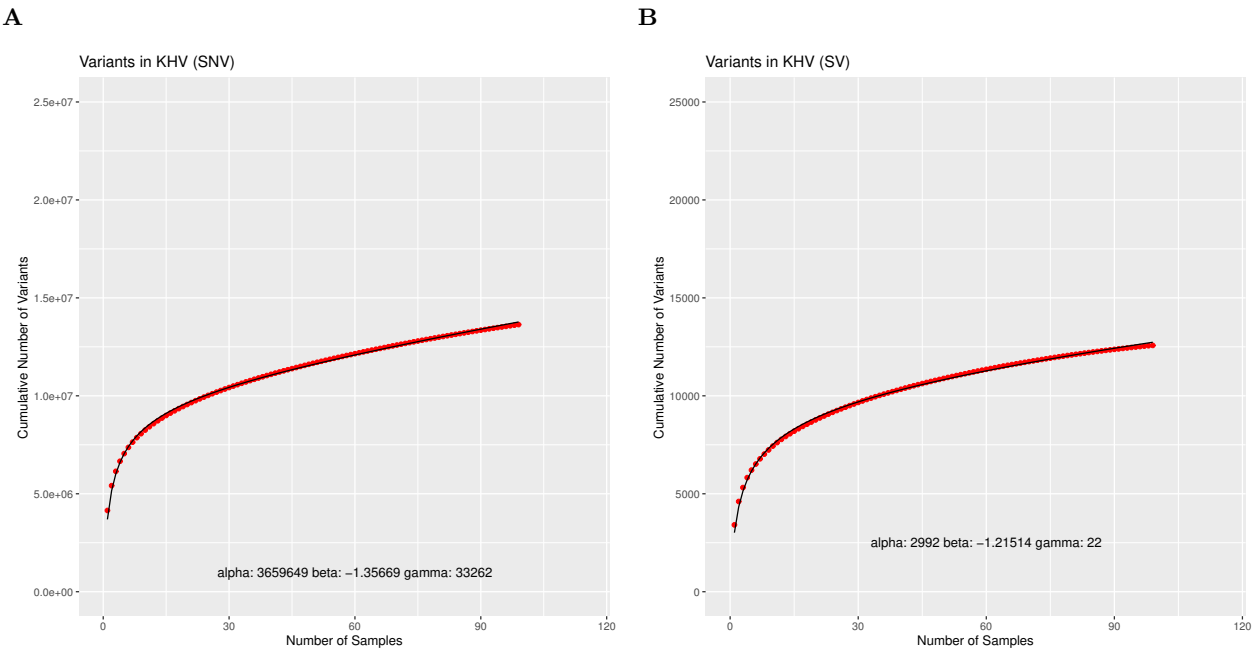

**Supplemental Figure S17:** SVCollector curves for the Iberian Population in Spain (IBS) population using **(A)** SNV data and **(B)** SV data.



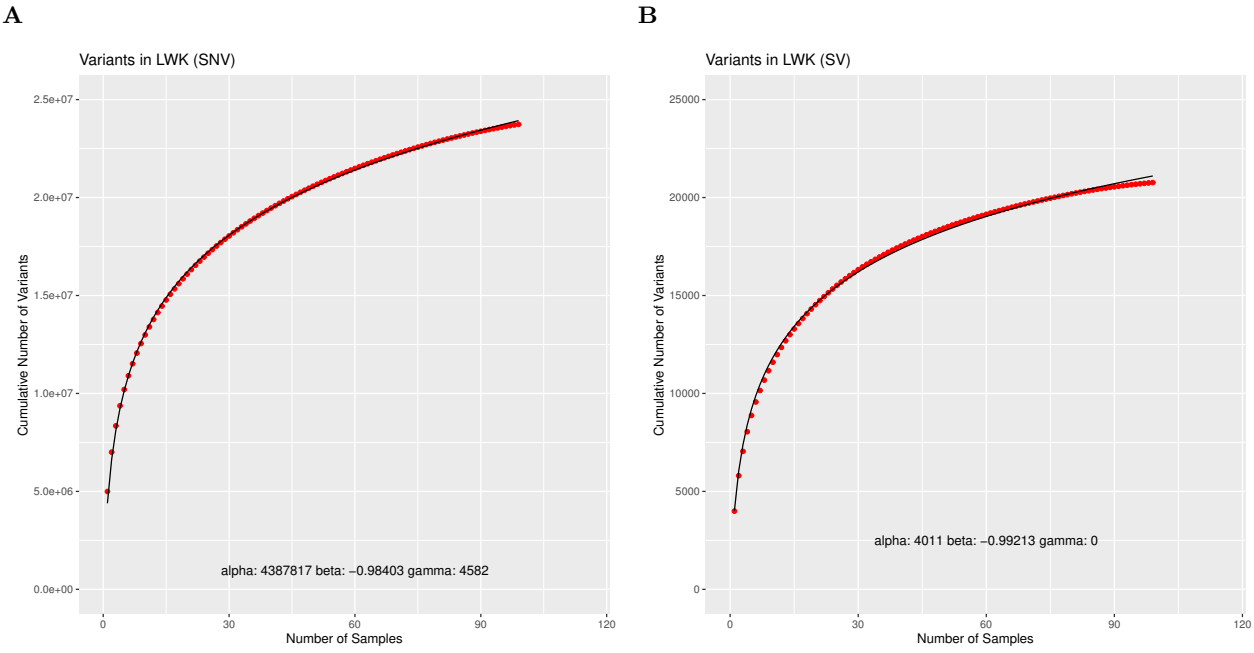

**Supplemental Figure S18:** SVCollector curves for the Indian Telugu from the UK (ITU) population using **(A)** SNV data and **(B)** SV data.



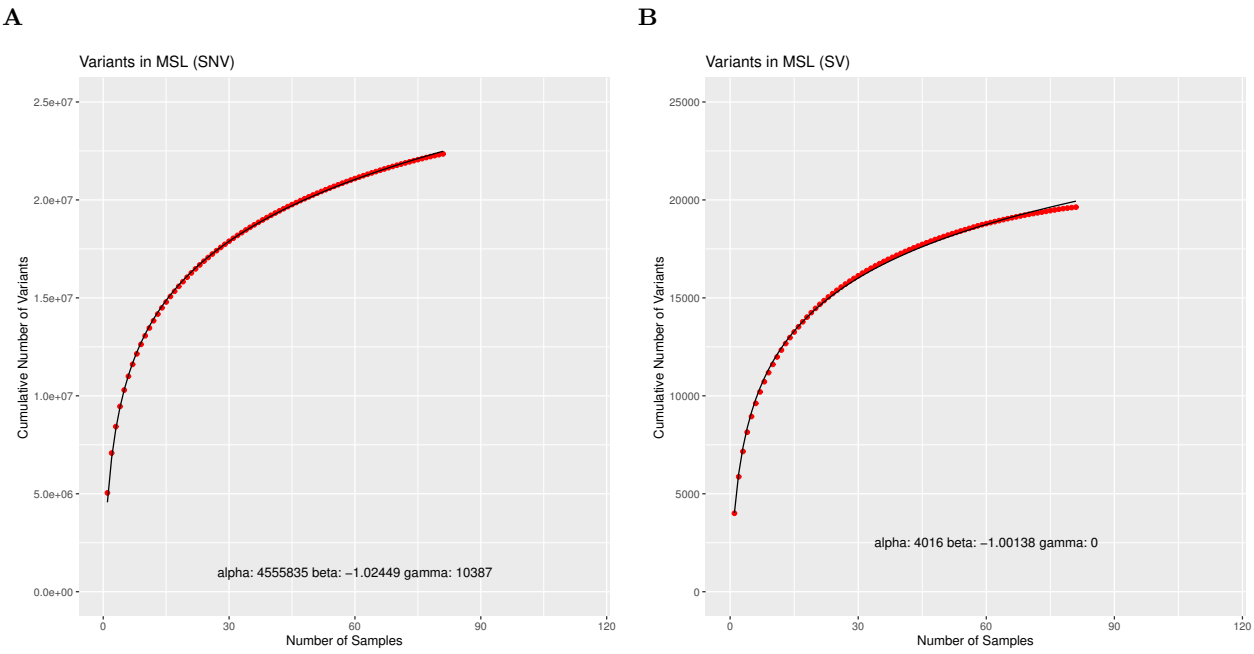

**Supplemental Figure S19:** SVCollector curves for the Japanese in Tokyo, Japan (JPT) population using **(A)** SNV data and **(B)** SV data.



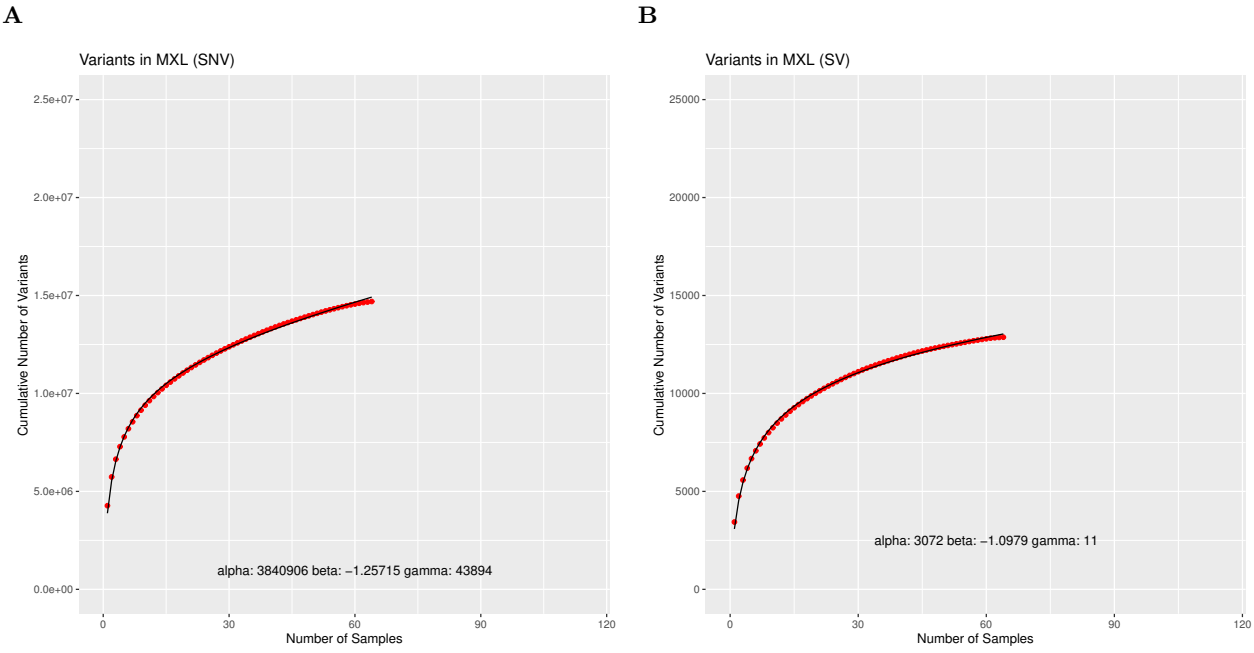

**Supplemental Figure S20:** SVCollector curves for the Kinh in Ho Chi Minh City, Vietnam (KHV) population using **(A)** SNV data and **(B)** SV data.



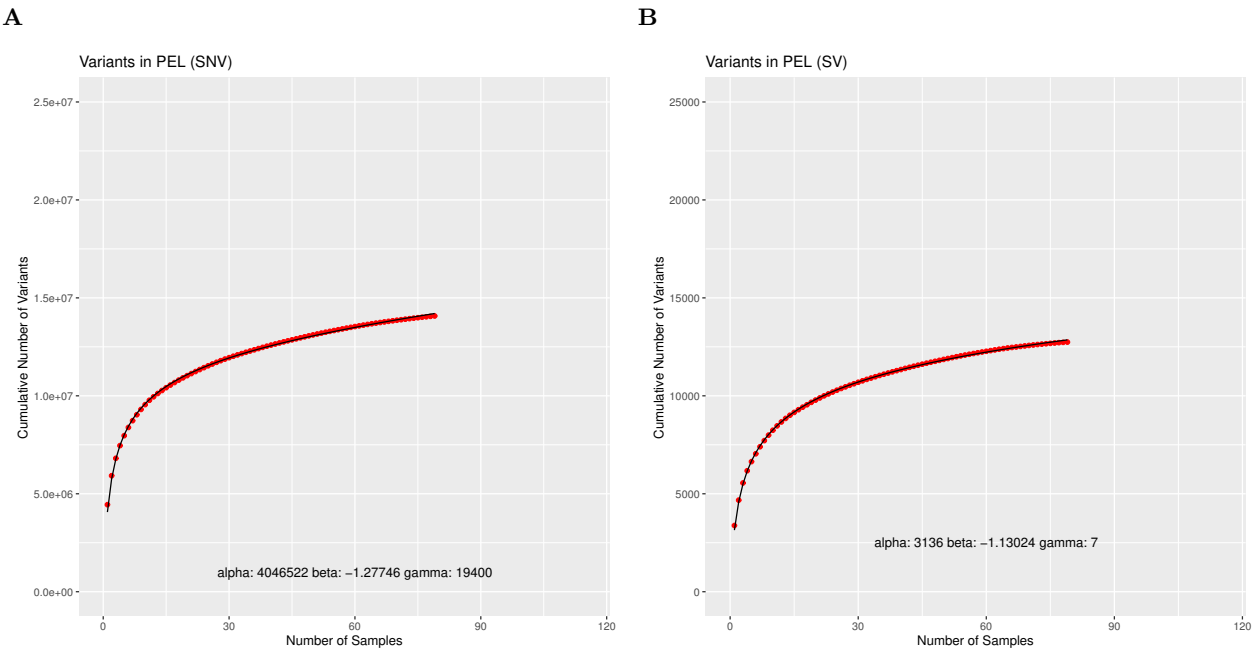

**Supplemental Figure S21:** SVCollector curves for the Luhya in Webuye, Kenya (LWK) population using **(A)** SNV data and **(B)** SV data.



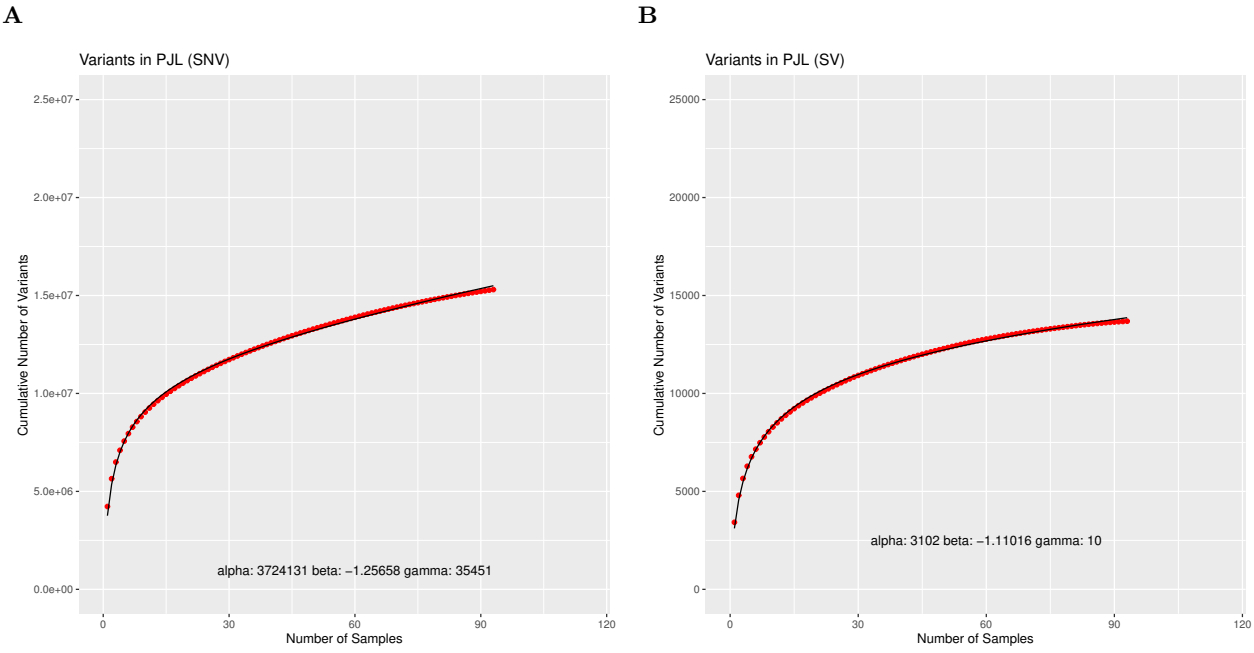

**Supplemental Figure S22:** SVCollector curves for the Mende in Sierra Leone (MSL) population using **(A)** SNV data and **(B)** SV data.



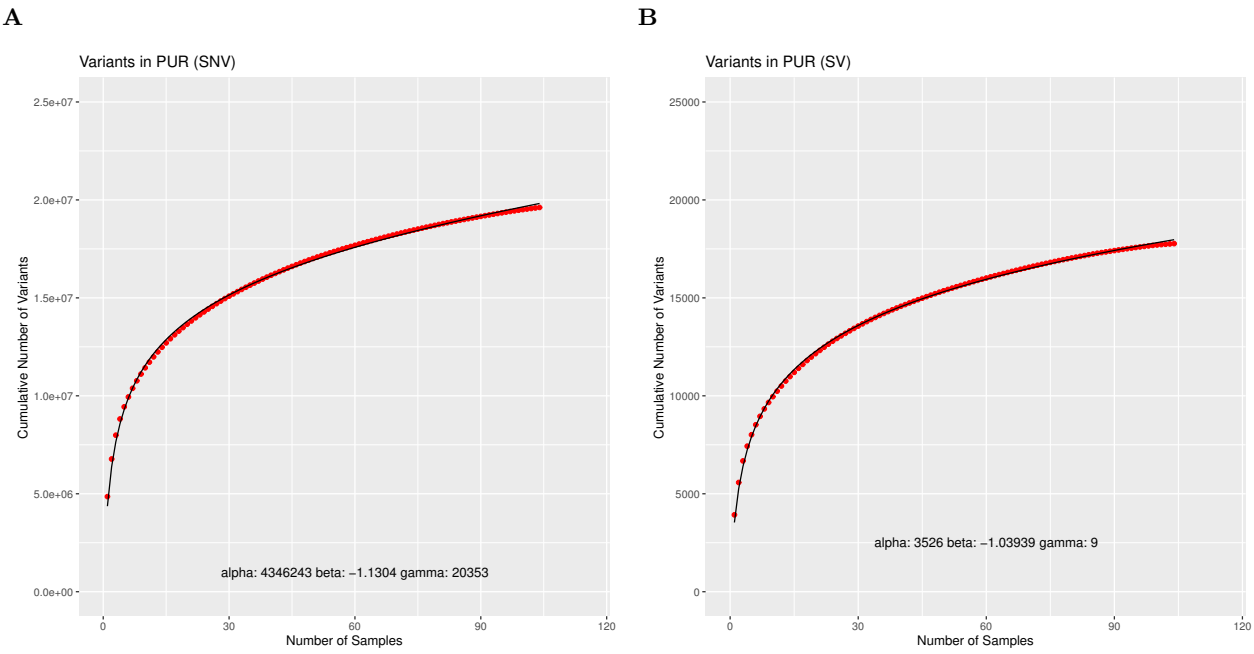

**Supplemental Figure S23:** SVCollector curves for the Mexican Ancestry from Los Angeles USA (MXL) population using **(A)** SNV data and **(B)** SV data.



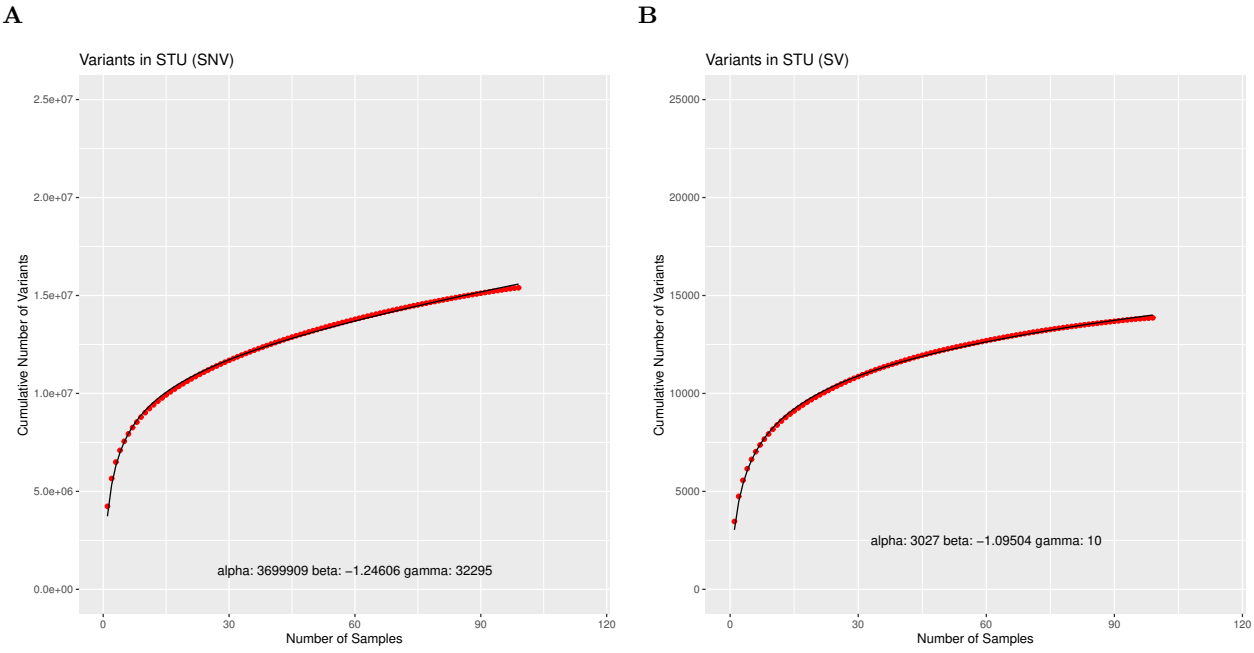

**Supplemental Figure S24:** SVCollector curves for the Peruvians from Lima, Peru (PEL) population using **(A)** SNV data and **(B)** SV data.



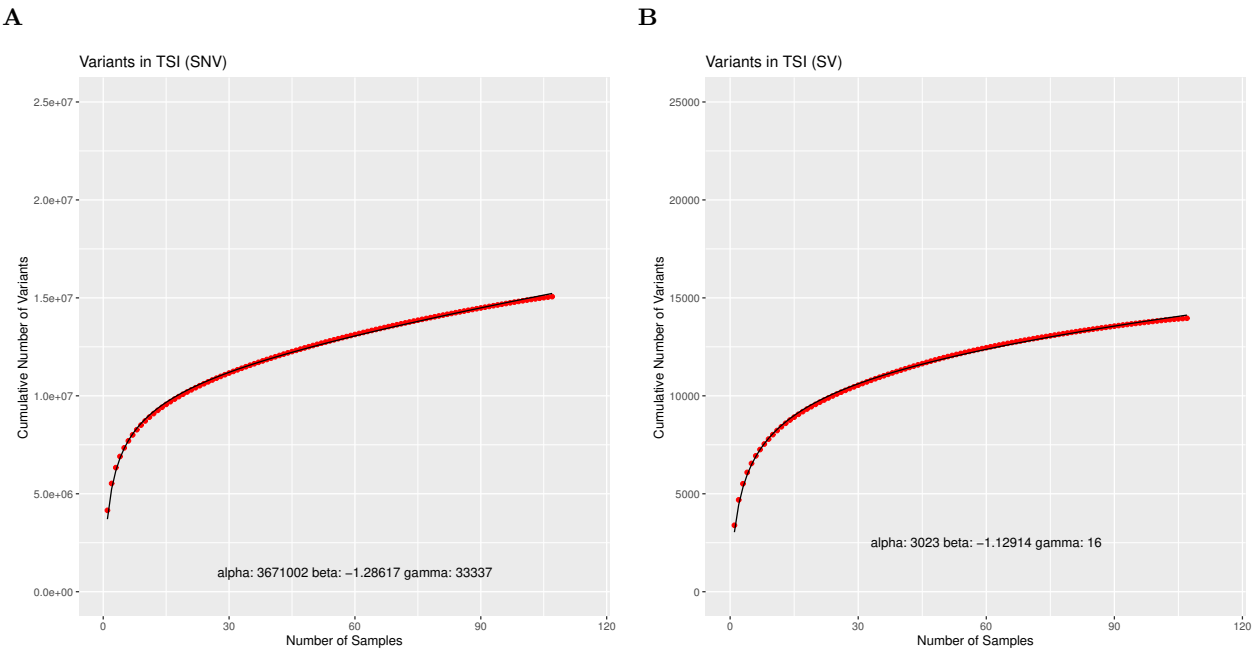

**Supplemental Figure S25:** SVCollector curves for the Punjabi from Lahore, Pakistan (PJL) population using **(A)** SNV data and **(B)** SV data.



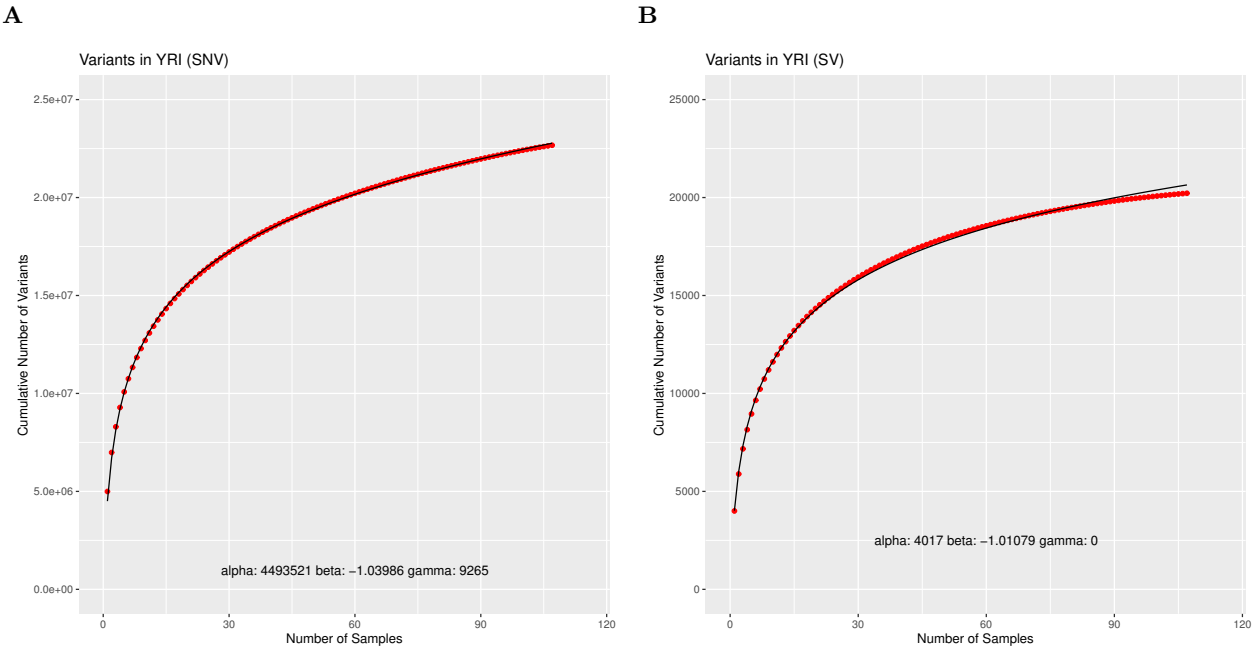

**Supplemental Figure S26:** SVCollector curves for the Puerto Ricans from Puerto Rico (PUR) population using **(A)** SNV data and **(B)** SV data.



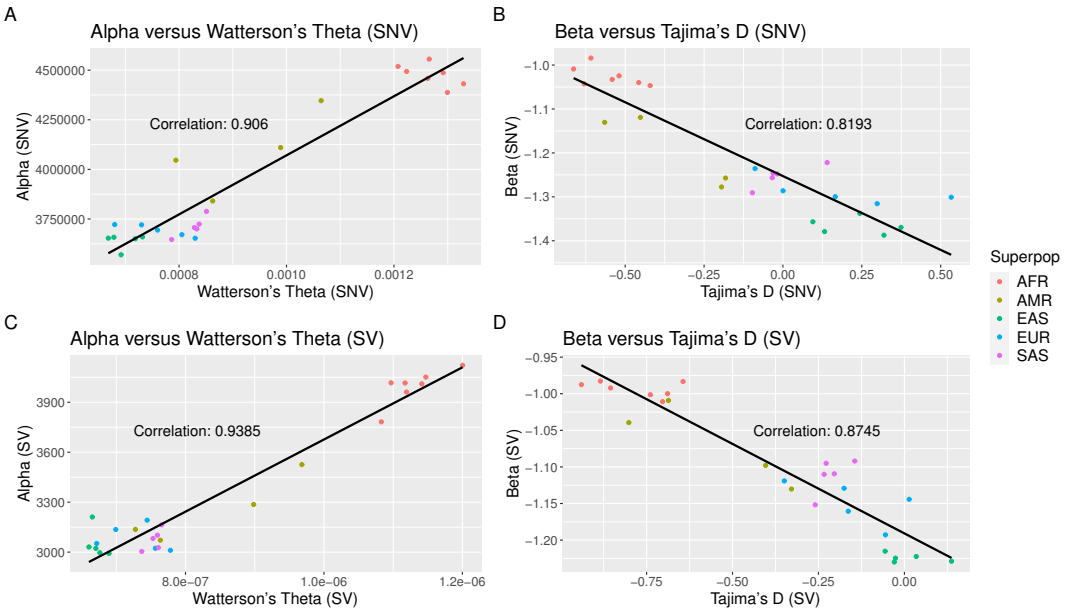

**Supplemental Figure S27:** SVCollector curves for the Sri Lankan Tamil from the UK (STU) population using **(A)** SNV data and **(B)** SV data.



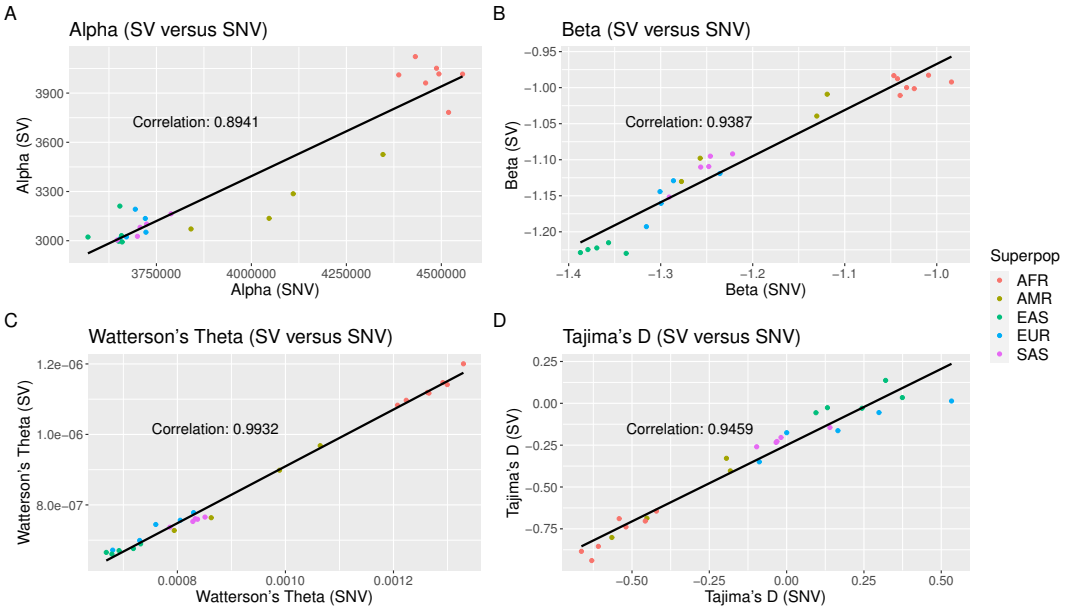

**Supplemental Figure S28:** SVCollector curves for the Toscani in Italia (TSI) population using **(A)** SNV data and **(B)** SV data.




**Supplemental Figure S29:** SVCollector curves for the Yoruba in Ibadan, Nigeria (YRI) population using **(A)** SNV data and **(B)** SV data.




**Supplemental Figure S30:** Linear regressions of power-law model parameters versus population diversity metrics. Each point represents one of the 26 subpopulations in the 1000 Genomes Project. The upper panels are calculated on SNV data and the lower panels are calculated on SV data.




**Supplemental Figure S31:** Linear regressions of power-law model parameters and population diversity metrics (SV versus SNV) for each of the 26 subpopulations.