Optimization of sequence coverage and its effects on population analyses

HudsonAlpha INSTITUTE FOR BIOTECHNOLOGY

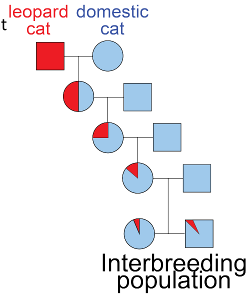
BioTrain

John M. DeLay; Chris Kaelin, PhD; Gopal Battu, PhD; Greg Barsh, MD, PhD

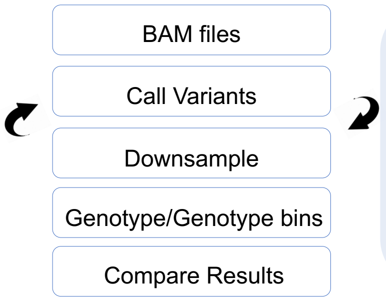
Coat color genetics has been a subject of intrigue since civilizations began artificially selecting for traits in livestock and domesticated animals. One interesting product of this breeder selection is the Bengal cat, a subspecies of the domestic cat that has gained popularity for its exotic look and high price. Bengal cats are hybridized descendants of Asian leopard cats and domestic cats; though they resemble domestic cats behaviorally and physiologically, they maintain the wild patterns of the Asian leopard cat. Systematic introgression of coat genes from backcrossing and interbreeding dilutes the parts of the genome inherited from the Asian leopard cat until there is a manageably small pool of candidate genes that may be responsible for the wild spotting in Bengal cats.

A sample population best contributes to structural studies when the pool is numerous, diverse, and reliable. Information is not free, and expensive sequences can force the investigator to choose between sample reliability and population breadth. In this study, we investigated the empirical effects of downsampling regarding common, whole-genome analyses used for identifying species-related marker genes in cats and found that **coverage as low as 0.2X is sufficient for deducing ancestry in regions throughout the genome in bins ranging from 10kbp to 200kpb.**

The top picture shows the striped pattern of the typical domestic cat. Below it is an example of the ornate, wild patterns of the Bengal cat. To the right of the coats, we see the pedigree tree demonstrating how Bengal breeders are able to retain the exotic coat color of an otherwise domestic cat.



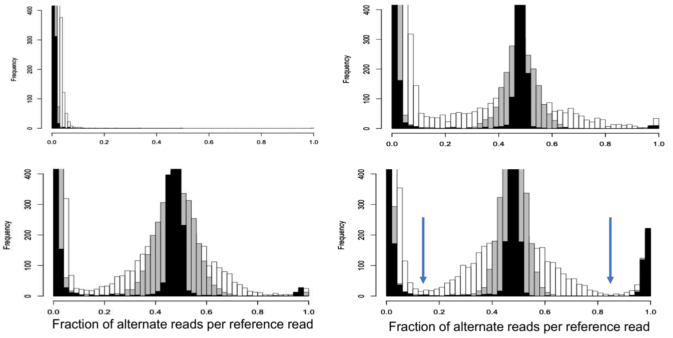
METHOD



To the left, we have the pipeline used for comparing the effects of low sequence coverage. We are given a set of aligned Bengal cat genomes. After calling variants and analyzing their diplotype by region, we progressively downsample and repeat the analysis process. In order to measure accuracy, we compare the individual locations to themselves across each downsampled iteration and plot the percentage of correct calls, taking the highest coverage files as truth. It is important to note that even the original file is subject to a degree of error as a result of the uncertainty in different processes through the project.

RESULTS

Distribution of alternate/reference fraction of 50 kb bins for one domestic cat and three Bengal cats



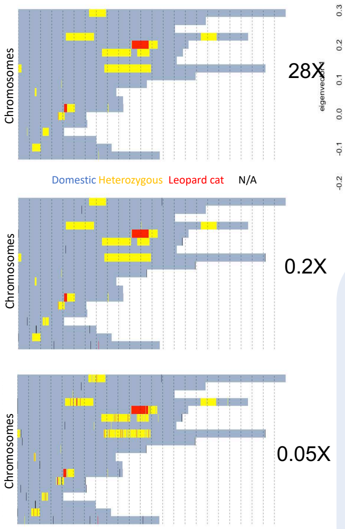
<u>Above</u> is a subset of the collection of histograms generated from the fraction of alternative read against reference reads in every bin across the genome. The black bars represent the number of bins that contain the respective allele frequency at the sample's original, high coverage. Those in grey are below 5X coverage, and in white, below 0.1X. For reference, Cat 1 is not a Bengal cat. Because there are few to no alternate reads in the marker subset, we see that the species-specific alleles are informative. In order to determine the fraction threshold for recoding the bin genotype, we studied over 300 of these histograms. Each graph contained significant troughs near 0.15 and 0.85 frequency (arrows in Cat 4), so we established those ratios as the ancestry cutoffs. After recoding the fractions into just homozygous reference (0), heterozygous (1), and homozygous alternate (2), each bin will suggest how the Asian Leopard alleles are inherited.

<u>Below</u> is the matrix used to compare correct calls. Accuracy of each coverage is determined by the fraction of each call type (0,1,2) retained through every position after each downsampling. We consider the original coverage as truth and scrutinize each of the subsequent files. In this hypothetical example, Coverage_0 is 100% correct and Coverage_2 is 33% correct.

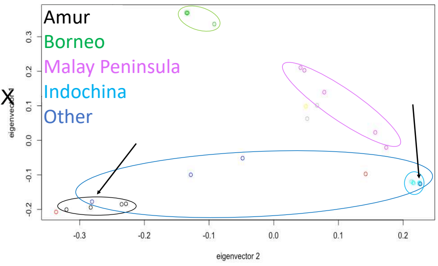
$$\begin{array}{c|cccc} & \text{Coverage}_0 & \text{Coverage}_1 & \text{Coverage}_2 & \cdots \\ \hline \text{Position}_0 & 1 & 1 & 0 & \\ \text{Position}_1 & 1 & 0 & 1 & \\ \text{Position}_2 & 1 & 1 & 0 & \\ \vdots & & & & \ddots \end{array}$$

RESULTS cont'd.

Diplotype plots of one Bengal cat at several depths of coverage in 50kb bins



Genetic structure of Asian leopard cats across their natural range

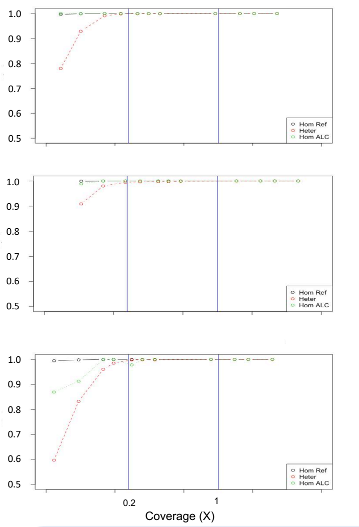


The figure on the <u>left</u> is a set of diplotype plots. They display the genome of the same cat but at different depths of coverage. Each small space in each chromosome is a bin (50kb in this subset), and each bin is color-coded according to the fraction of alternate reads per reference read. The alleles we considered were selected from a previous study because they were species-informative (one way in domestic and another in Asian leopard). Recall that the sequence is diploid, so blue is homozygous domestic, yellow is heterozygous, and red is homozygous Asian Leopard – black represents a region with no alignment reads. By the human eye, we see that there is a larger disparity of correct calls between 0.2X and 0.05X than between .2X and 28X.

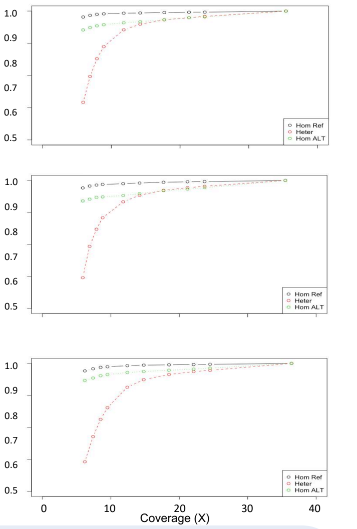
<u>Above</u> is a principal component analysis of the individual marker variants in 27 Asian leopard cats. The arrows point to entries from two sets of variants that came from the exact same cat – their only difference is their coverage: 3X (left) and 38X (right). Their distribution, compared to other encircled subgroups, exhibits the potential impact that low-coverage samples might have on population studies.

At low coverage, we can gather nearly the same information from a diplotype plot. In analyses that rely on individual variants such as principal component analysis, we see the risk of falsely categorizing a sample due to the variant error produced by low coverage.

Fraction of correct 50kb bin calls for Cats 2,3,4 vs. coverage



Fraction of correct variant calls for three zebras vs. coverage



Above are two sets of data. The <u>left</u> maps the error trend calculated in each Bengal cat during the downsampling process. Line colors identify the different genotype calls (Black: homozygous domestic; Red: heterozygous; Green: homozygous leopard cat). Note that the coverage axis is logarithmic and that as coverage decreases, error becomes more sensitive to coverage.

Consider an analogous study on the right. These charts plot the error trends of genotype calls at individual loci. Compared to the zebra data, we see that genotype bins are far less sensitive to depth of coverage. Since ancestry is usually inherited in regions larger than 50kb, we can conclude that larger bins will be as or more accurate than this subset of information. Even at very high coverage, genotyping software cannot deliver error-free variants, so we see that affordable, lower-coverage sequences provide valuable information for population-scale studies that investigate large regions.

<u>Tools:</u> VCFtools, Platypus, FreeBayes, R, SNPRelate