Detailed Alignment Pipeline

ROUND 1

1. Mapping

- *BWA* under default (strict) settings
- Pass all unmapped reads to *Stampy*

2. Filter reads with mapping quality <20

3. Mark optical duplicates

4. Realign around indels using GATK Indel Realigner

5. Call SNPs using GATK Unified Genotyper

- Minimum alignment base quality of 31
- Filter SNPs to N with <0.75 proportion of reads supporting SNP

6. Call indels using GATK Unified Genotyper

- Minimum of 3 reads possessing indel to call it
- Minimum of 0.51 proportion of reads must agree to call indel

7. Modify the reference to be used for this genome in round 2

- Insert SNPs and INDELS called above into reference genome

ROUND 2

8. Map to modified reference

- *BWA* under default (strict) settings
- Pass all unmapped reads to *Stampy*

9. Filter reads with mapping quality <20

10. Mark optical duplicates

11. Realign around indels using GATK Indel Realigner

12. Call all sites using GATK Unified Genotyper (***with AllSites option)***

- Minimum alignment base quality of 75 for haploid-embryo genomes, minimum quality of 32 for diploid genomes

13. Shift SNP and indel coordinates back to those of the original reference genome

- Custom programs available upon request (jlack@wisc.edu)

14. Filter all sites to N within 3 bases of indels

15. Heterozygosity filtering of inbred line genomes (full intervals masked to N)

- 100 kb windows, sliding 5 kb
- Begin excluding windows where the proportion of heterozygous sites exceeds $\pi/5$ for that window, and continue excluding windows in both directions until the proportion of sites that are heterozygous is below $\pi/20$ for that window. For cosmopolitan genomes, π was calculated using genomes from the French (FR) population, and for African genomes π was calculated using the Rwandan (RG) population sample of genomes.

16. Pseudo-heterozygosity filtering of haploid embryo genomes (intervals masked)

- For haploid embryo genomes, copy number and structural variation can result in mismapping, producing regions of "pseudo-heterozygosity". Larger intervals were masked to N using a similar proces as used for true heterozygosity:
- 100 kb windows, sliding 5 kb
- Begin excluding windows where the proportion of sites with <75% of reads supporting the called nucleotide exceeds $\pi/5$ for that window, and continue excluding windows in both directions until the proportion of sites with <75% of reads supporting the called nucleotide is below $\pi/20$ for that window. For cosmopolitan genomes, π was calculated using genomes from the French (FR) population, and for African genomes π was calculated using the Rwandan (RG) population sample of genomes.

17. Identity-by-descent (IBD) filtering

- 500 kb windows, sliding 100 kb
- Windows with pairwise distance below 0.05% were considered IBD
- Only within-population comparisons were considered
- Relatedness IBD was marked for optional filtering (in one of a pair of related genomes) when summed genome-wide IBD tracts located occurred outside of "recurrent IBD regions" exceeded 5% of the bases called in both genomes. Recurrent IBD regions identified for each data group were as follows:

DGRP

chrX	1	2500000	telomere
chrX	20600001	22422827	centromere
chr2L	1	800000	telomere
chr2L	10900001	12000000	other recurrent IBD
chr2L	17700001	23011544	centromere
chr2R	1	5800000	centromere
chr2R	20000001	21146708	telomere
chr3L	19400001	24543557	centromere
chr3R	1	4900000	centromere
chr3R	5500000	9700000	other recurrent IBD
chr3R	15000001	27905053	other recurrent IBD / telomere

DPGP3

chrX 1 600000 telomere
chrX 13300001 14500000 other recurrent IBD
chrX 17700001 18000000 other recurrent IBD
chrX 19300001 19800000 other recurrent IBD
chr2L 1 700000 telomere
chr2L 17500001 23011544 centromere
chr2R 1 4300000 centromere
chr2R 7100001 7500000 other recurrent IBD
chr2R 15200001 16200000 other recurrent IBD
chr2R 20500001 21146708 telomere
chr3L 16400001 24543557 centromere
chr3R 1 9700000 centromere
chr3R 19600001 20200000 other recurrent IBD
chr3R 27400001 27905053 telomere

DPGP2/Pool

chrX 1 1800000 telomere
chrX 13200001 16400000 other recurrent IBD
chrX 17500001 19700000 other recurrent IBD
chrX 21300001 22422827 centromere
chr2L 1 800000 telomere
chr2L 5200001 7400000 other recurrent IBD involving a small group
chr2L 14400001 16400000 near centromere
chr2L 17600001 23011554 centromere
chr2R 1 5300000 centromere
chr2R 15400001 16300000 other recurrent IBD
chr2R 20200001 21146708 telomere
chr3L 14300001 24543557 centromere
chr3R 1 9800000 centromere
chr3R 14800001 23000000 intermittent IBD involving a small group
chr3R 23000001 27905053 telomere

18. Admixture filtering

In order to enable the analysis of genetic variation from the species' African ancestral range, we enable the masking of genetic variation in sub-Saharan African genomes that is inferred to have recent introgressed from non-sub-Saharan ("cosmopolitan") populations. The basic method of identifying ancestry along each chromosome is the same as described by Pool et al. 2012 PLoS Genetics. The only differences are as follows:

- All 27 Rwanda RG genomes can now be used in the sub-Saharan reference panel. Also, in part because window lengths scale based on Rwanda diversity, they tend to be a bit smaller now, since slightly more non-singleton SNPs can be called.
- Homozygous sections of Egypt genomes were added to the cosmpolitan reference panel. All 9 France genomes were also used.
- Chromosome arms carrying inversions were excluded from reference panels.