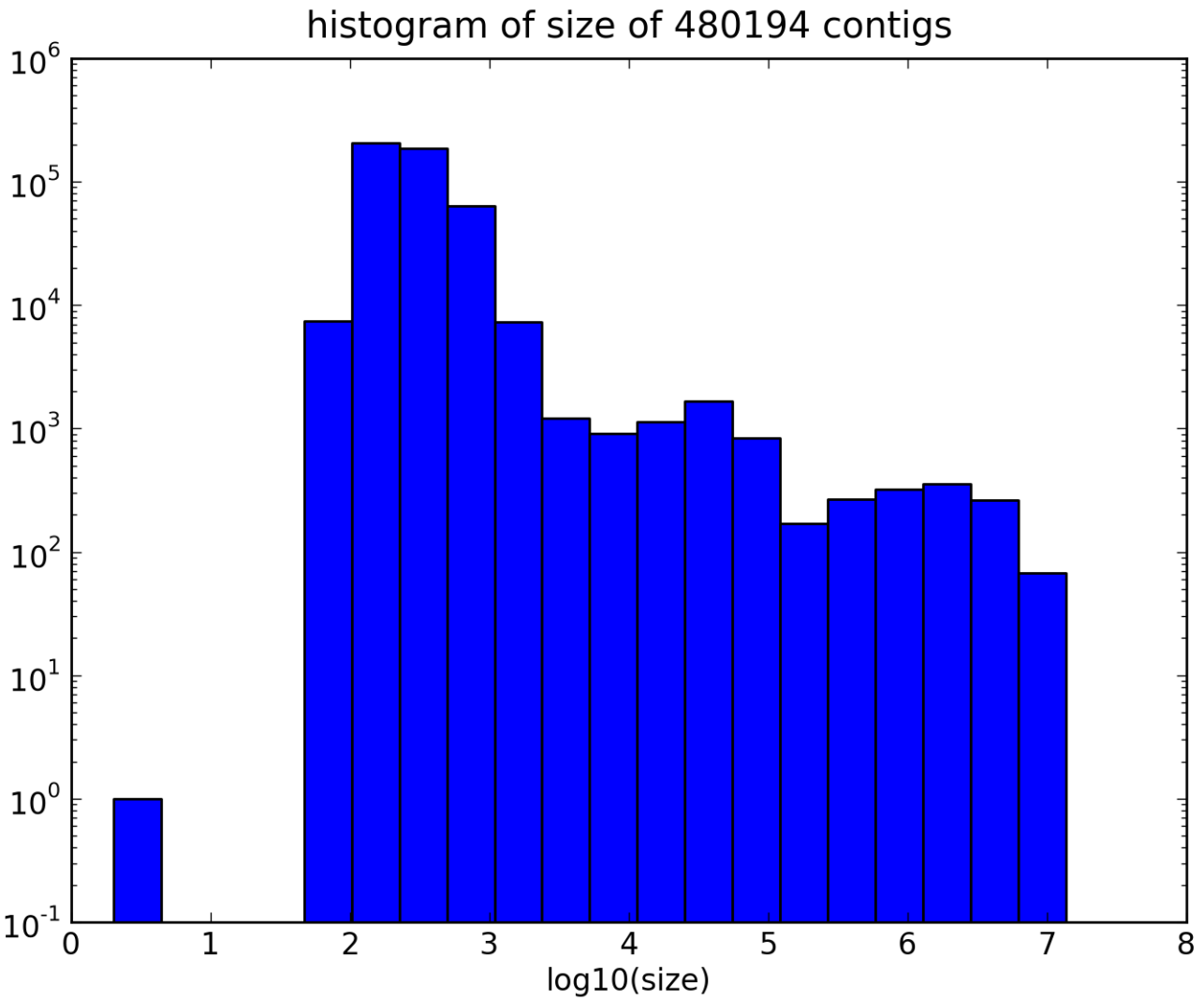
Population Genetics Analysis of Vervets

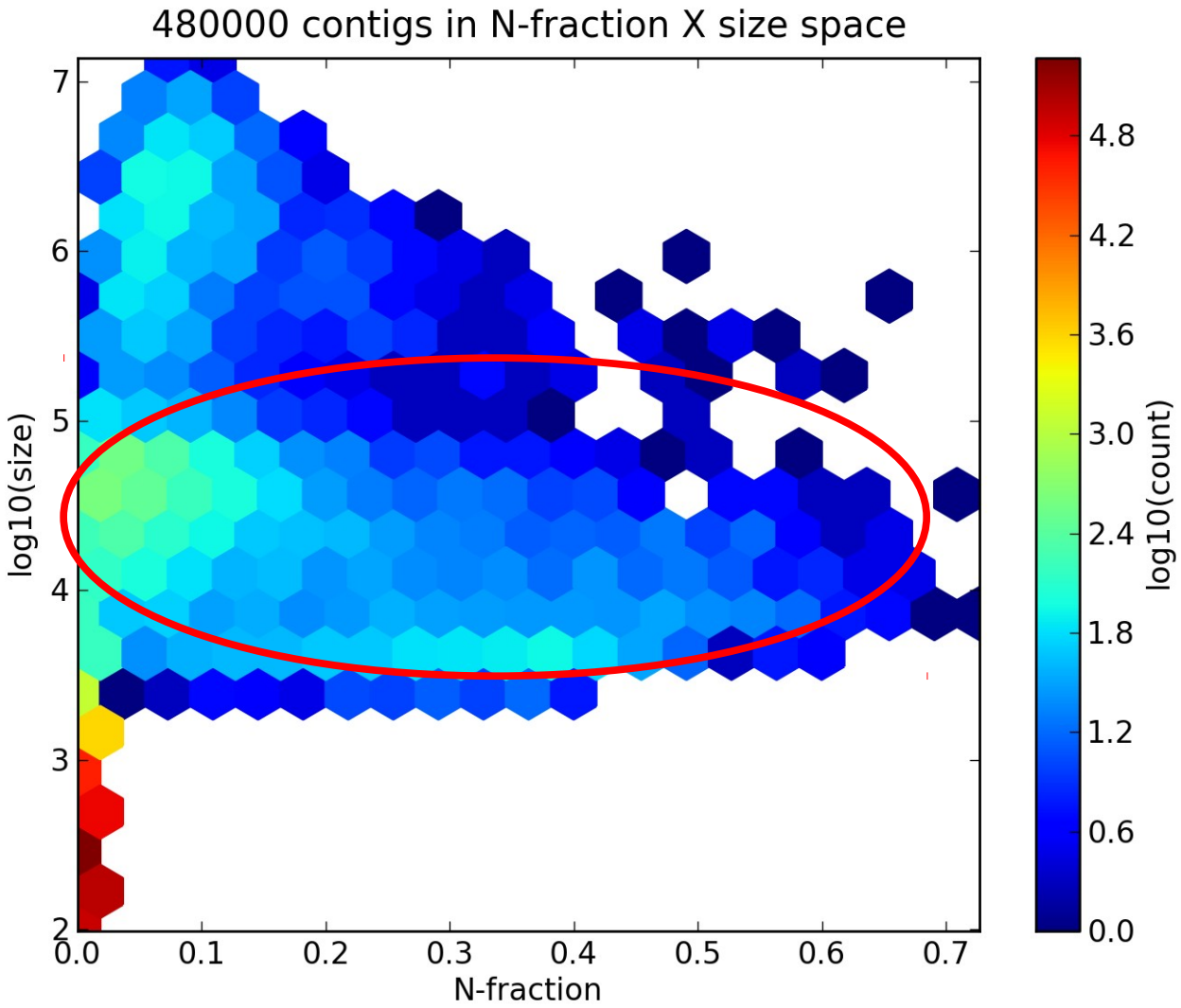
Data: Reference

- From VRC
- Sequenced by 454 at 8X (and more) and Illumina 100bp PE at ~10X.
- Current assembly is based on 454 reads.
  - ~480K contigs in total. Size from 2bp to 13.4Mb.
  - ~7K (size>=2Kbp) used in alignment.
  - Variants called over top 156 contigs (~1Gb in total size)

Size histogram of ~480K Supercontigs



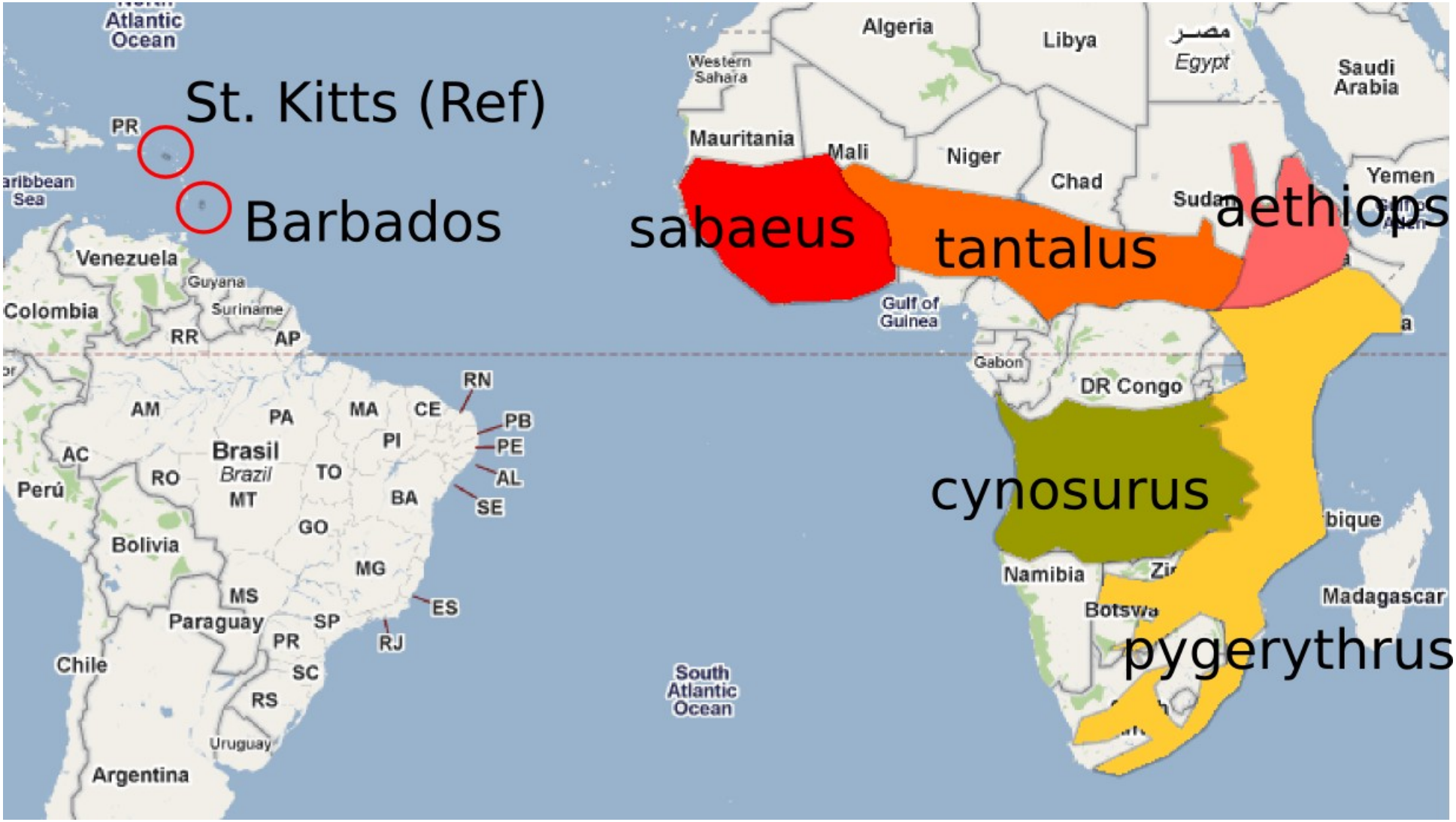
Median-sized contigs (10kb-100kb) have lots of missing bases.



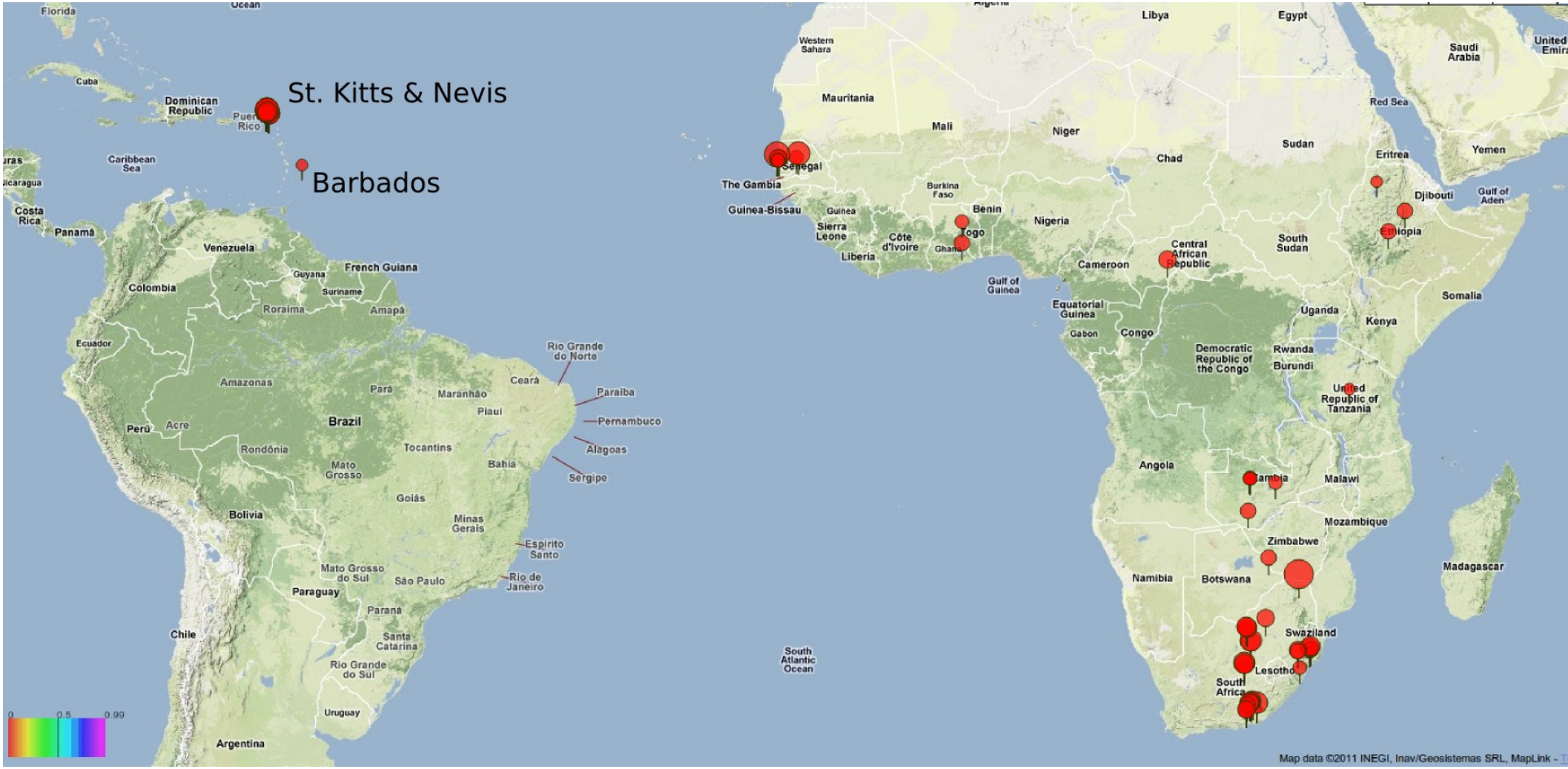
Data: Population Sequencing

- 116 out of ~1000 VRC monkeys
  - Most are sequenced by Solexa PE at 4-5X.
  - Four are sequenced at 30-40X.
  - Due to concern about the HWE violation,
    - 15 most distant ones (>=4 meiotic distance) are selected into an additional set.
- 5 St. Kitts + 5 Nevis
- 1 Barbados
- 5 African subspecies

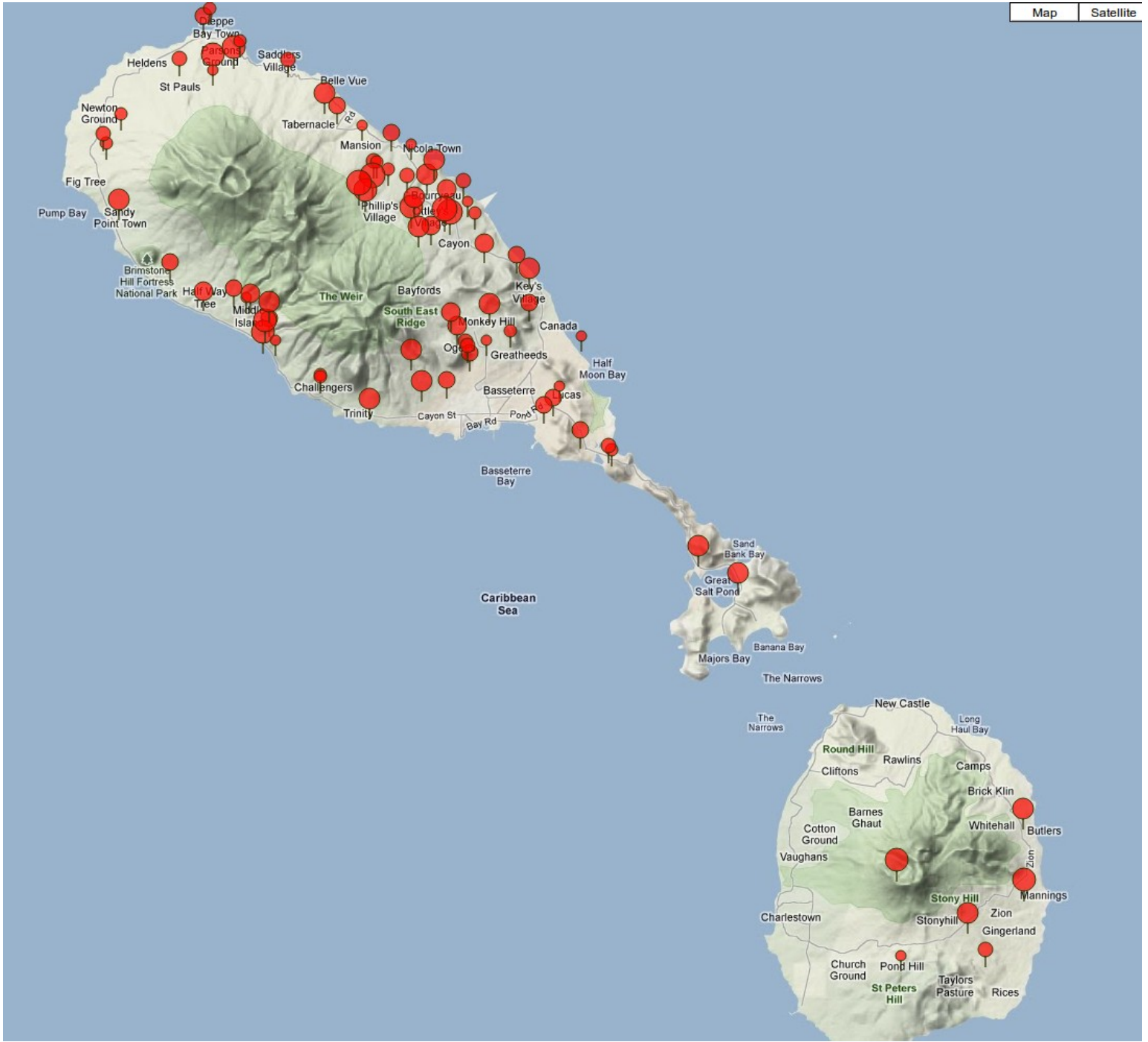
Geographic Map



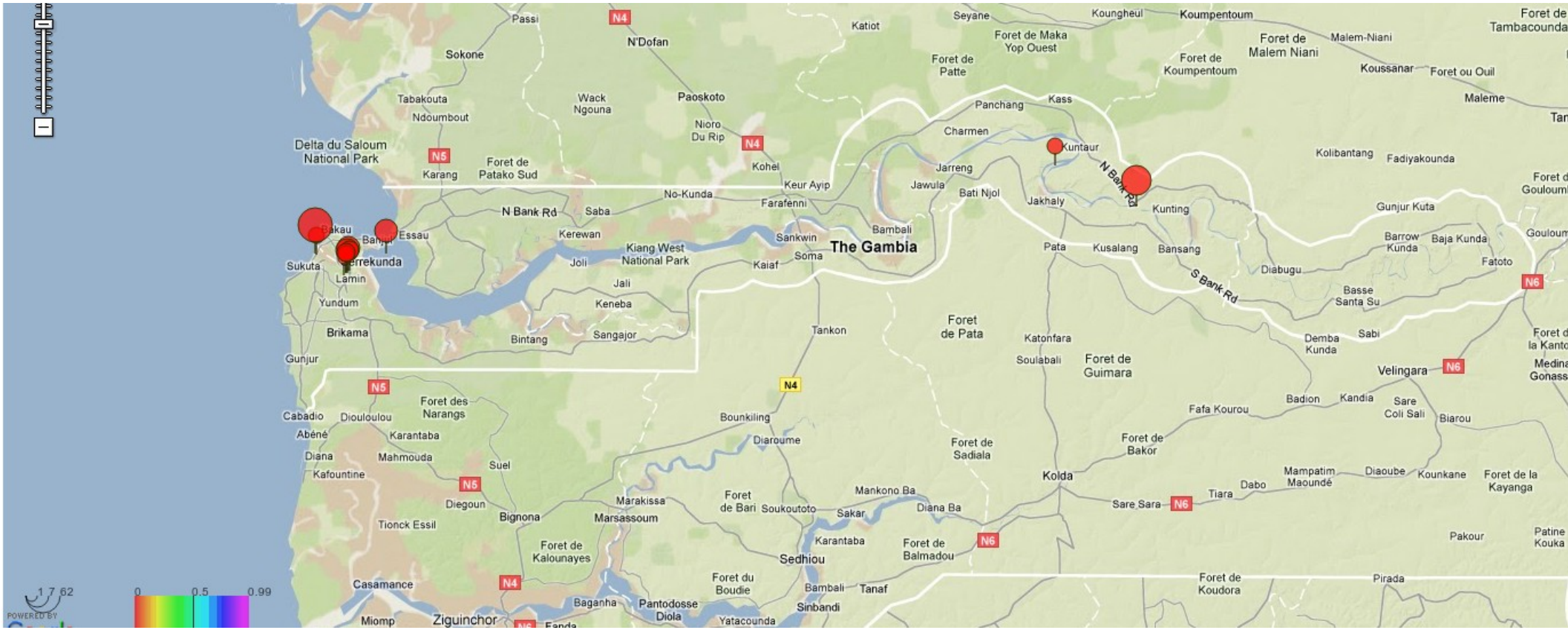
Vervet World Wide Collection



Collection at St. Kitts & Nevis



Collection at Gambia



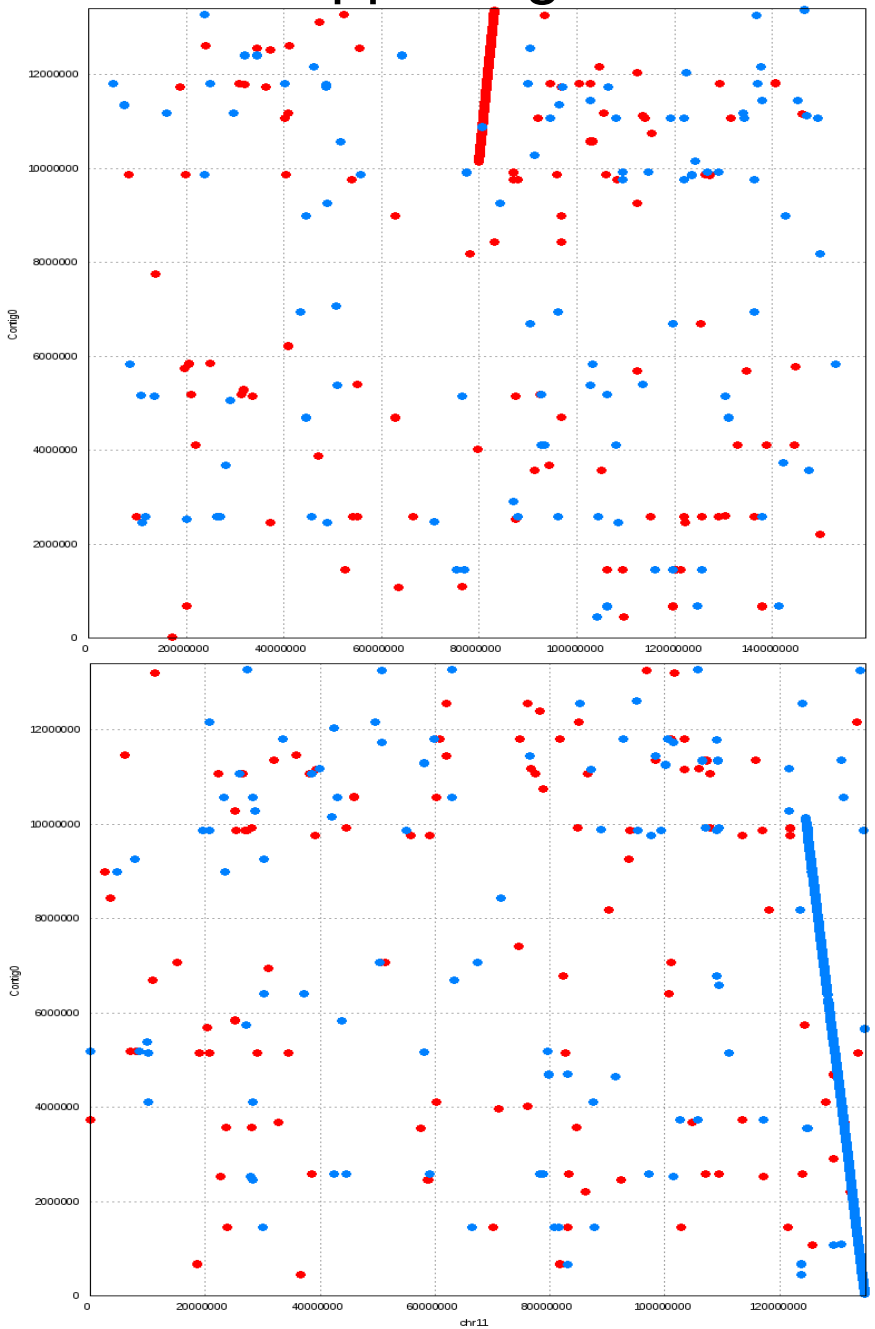
Outline

- Vervet-Human and Vervet-Macaque Synteny
- SNP variant pipeline and QC through trios
- Relationship of Carribean monkeys to African subspecies
- St. Kitts v.s. Nevis
- VRC v.s. St. Kitts & Nevis
- Outlook

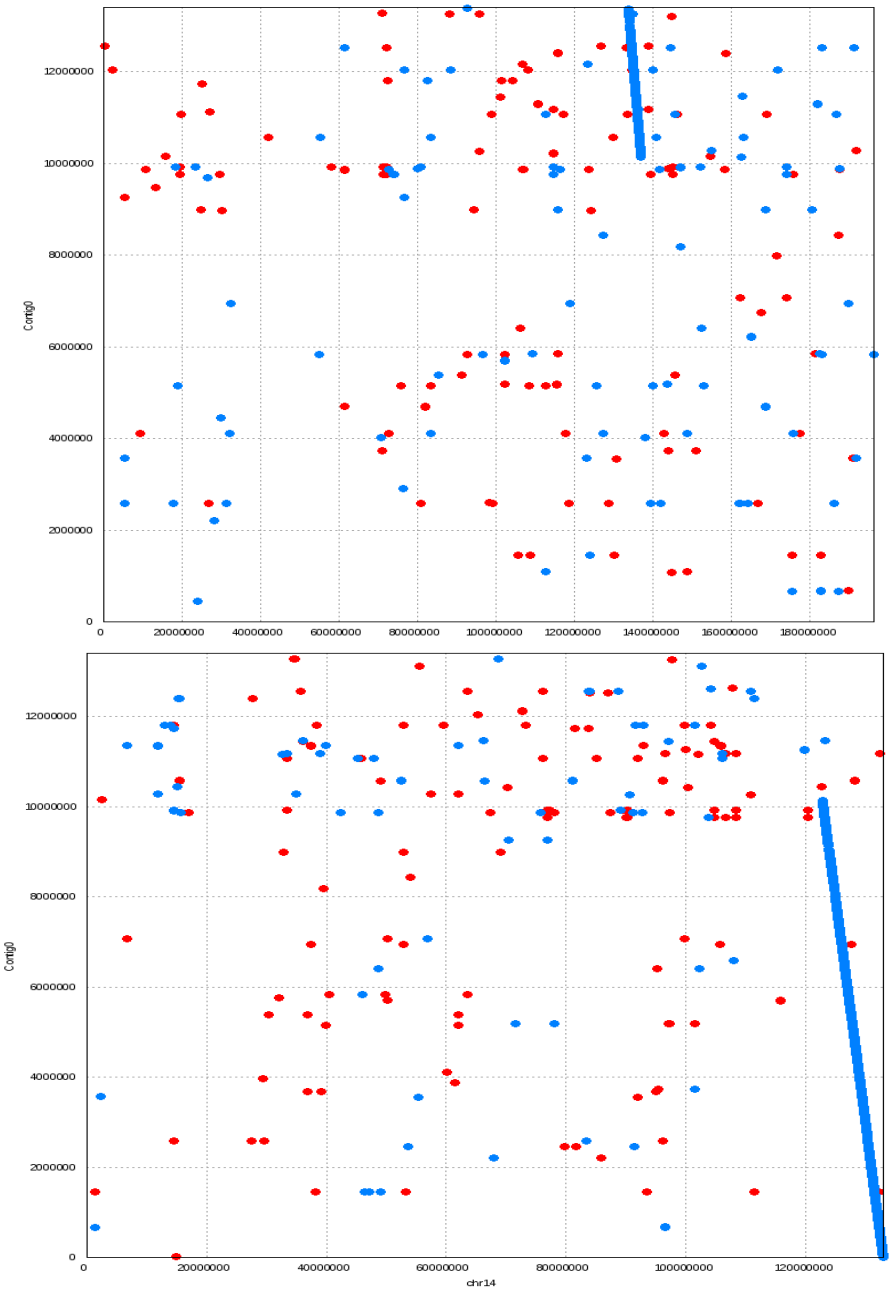
Synteny

- Top 156 contigs from vervet reference.
- hg19 for human reference.
- Latest Macaque genome.
- Run mummer http://mummer.sourceforge.net/.

A translocaton happens in MRCA of human & macaque?
Supporting human & macaque closer. Contig0.

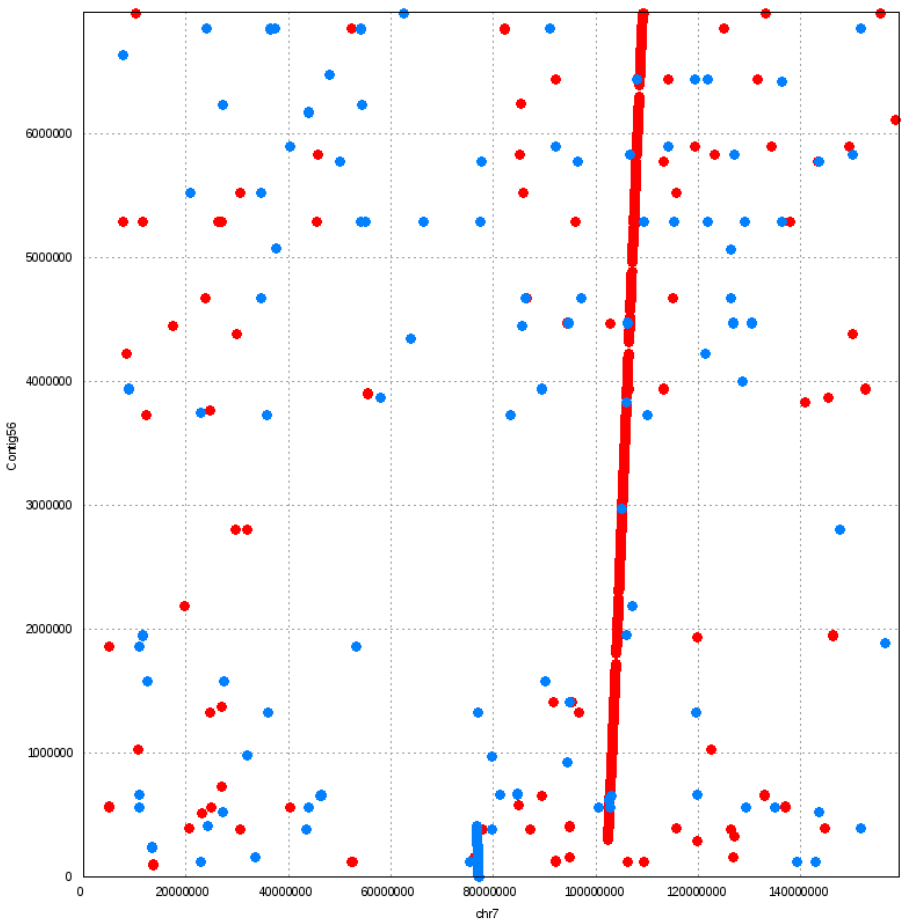


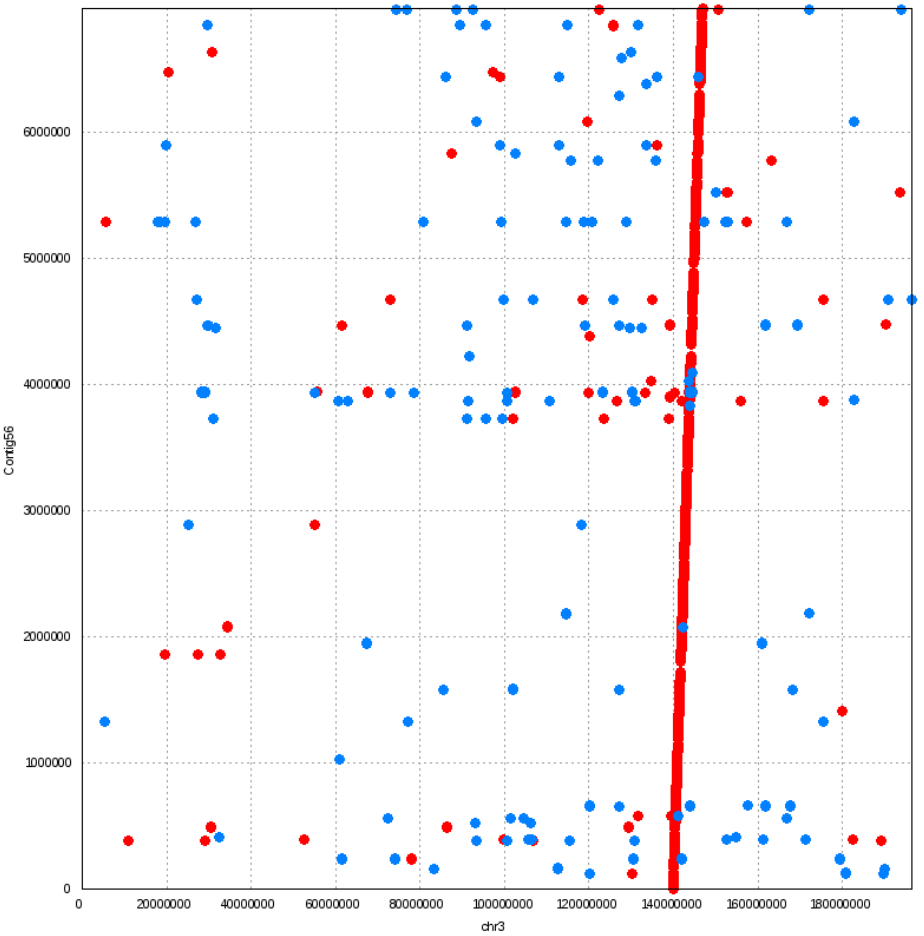
human chr7/chr11



macaque chr3/chr14

A translocation in human, aided by a duplication, supporting macaque & vervet closer. Contig56

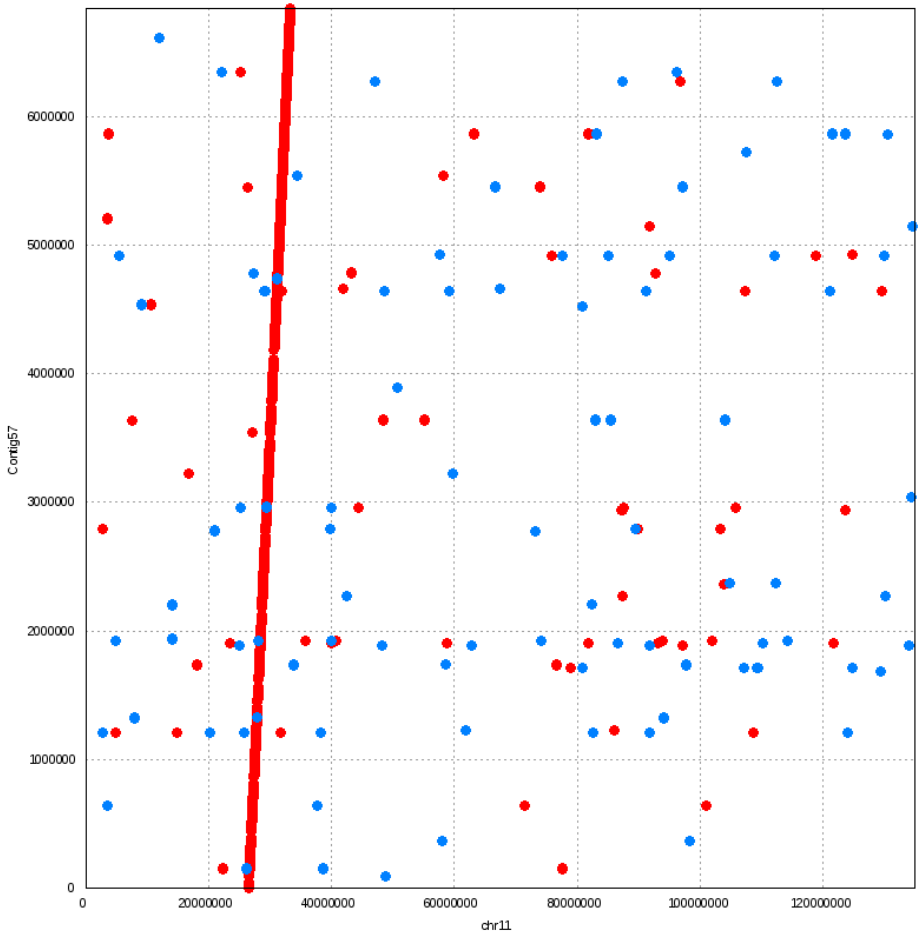


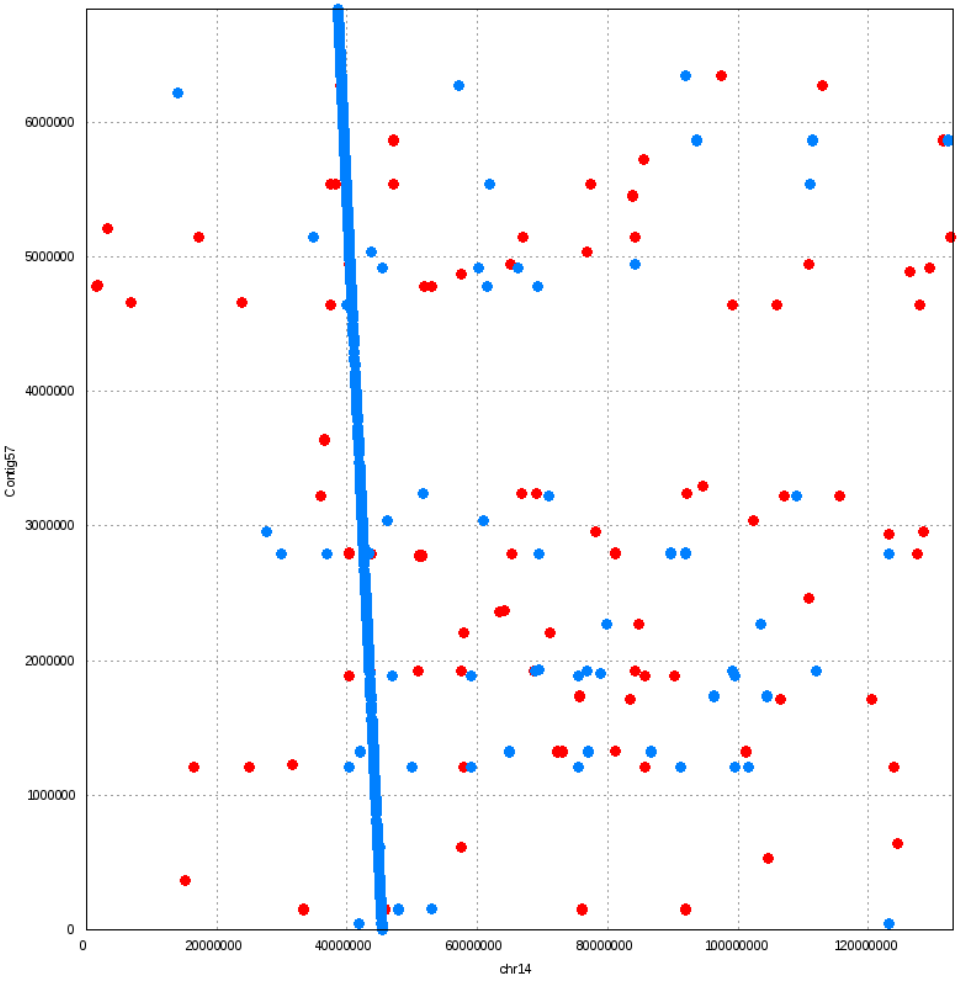


human chr7

macaque chr3

Contig57. No large-scale structural event.





human chr11

macaque chr14

Support for Different Trees by Synteny

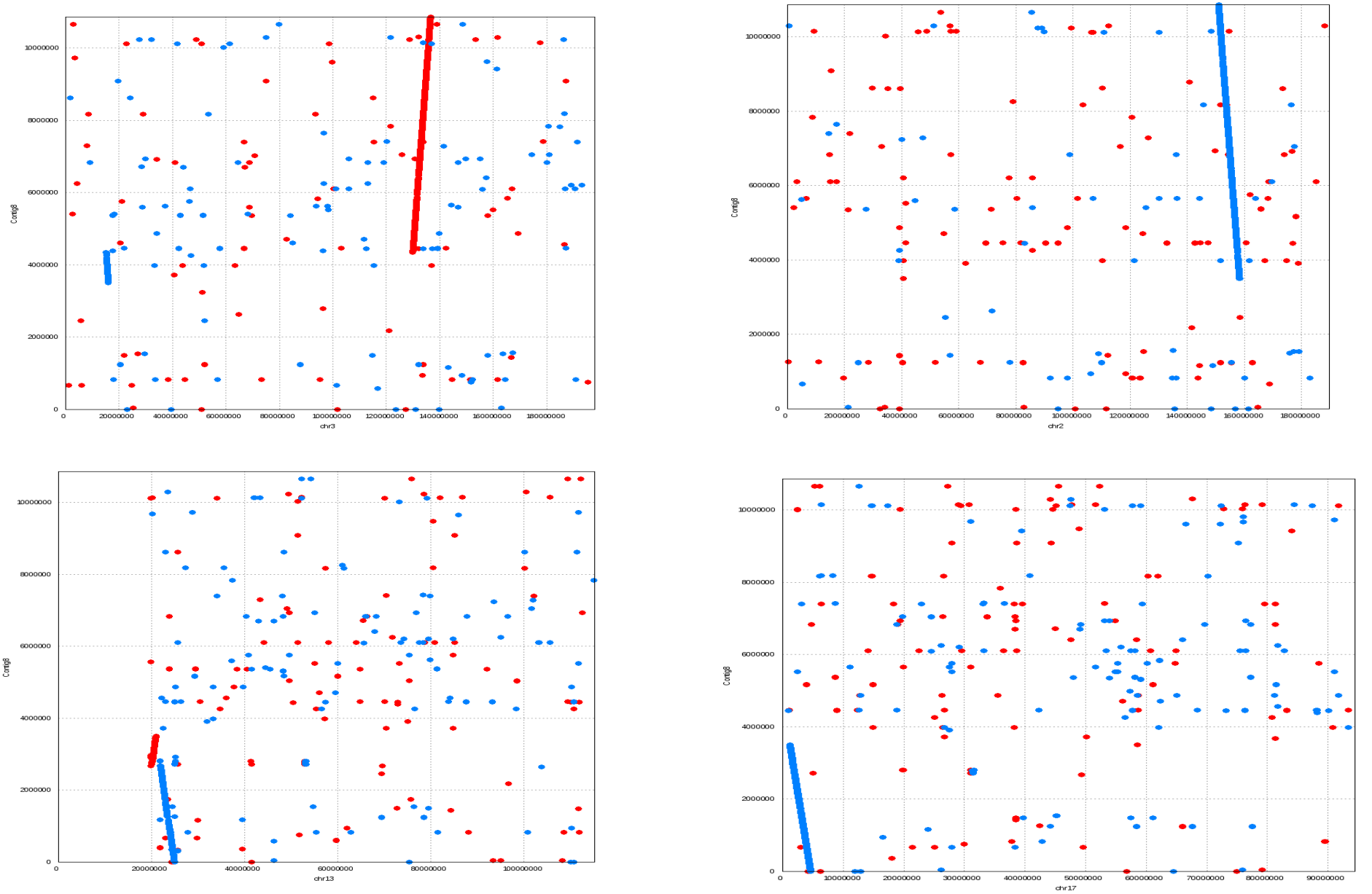
Eyeballing based on top 11 contigs (11 because 156 is too many for eyeballing).

Synteny support	**Number of contigs**
human & macaque closer	7.5
macaque & vervet closer	1.5
human & vervet closer	1
equally distant	1

But alignment to macaque genome shows fewer mismatches than to human reference.

About that 0.5, next slide.

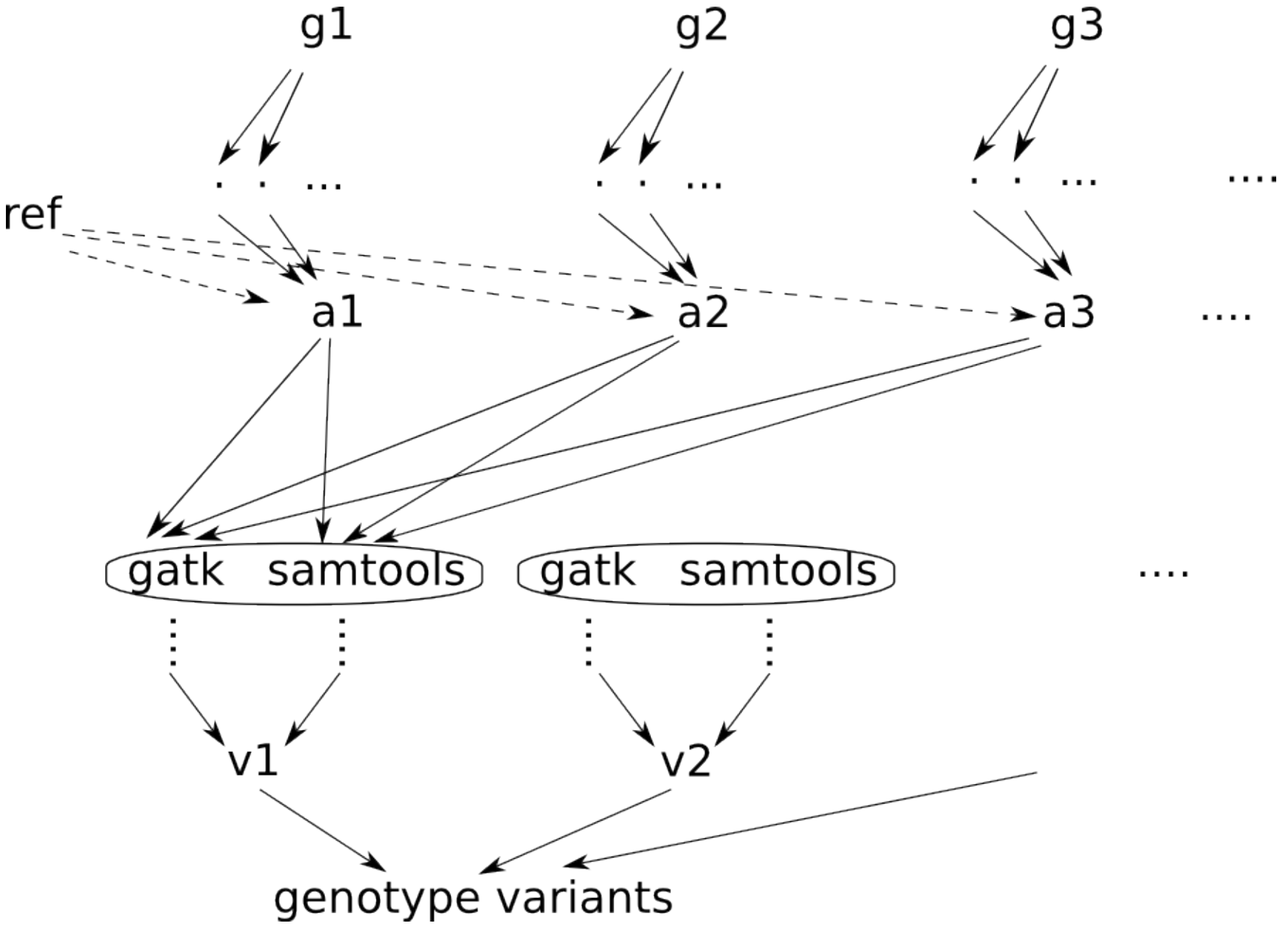
1st. One translocation happens in MRCA of human & macaue.
Then, one translocation + two inversions happen in human.



SNP Pipeline

- Use top 7559 contigs (size>=2kb) as reference.
- Run alignment for every individual's reads independently.
- Select a population of individuals (i.e. St. Kitts).
- Run GATK & samtools on alignments from the whole population (2MB interval by interval).
- Take the intersection of GATK & samtools calls.

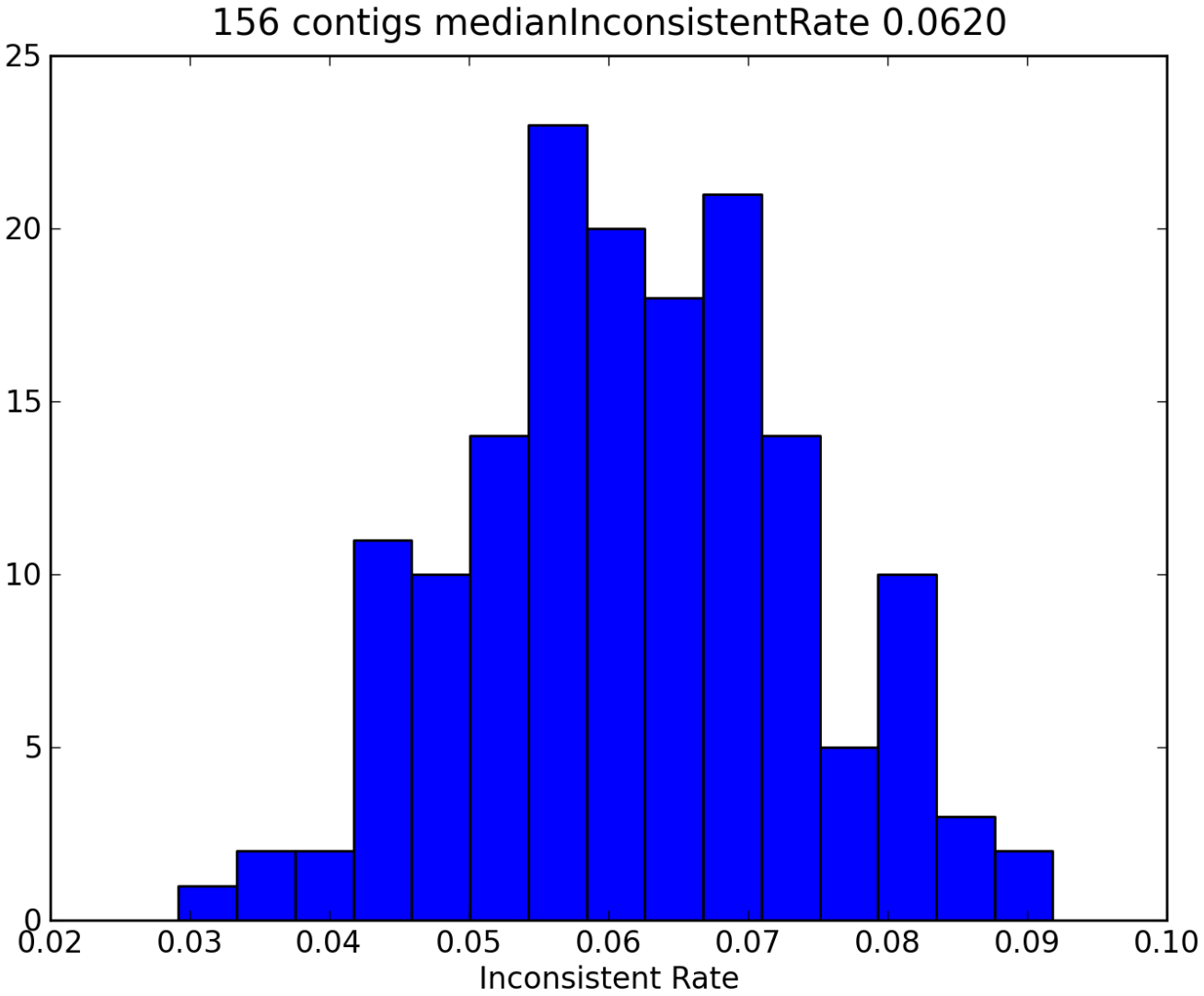
Diagram for one population.



SNP QC based on trio inconsistency

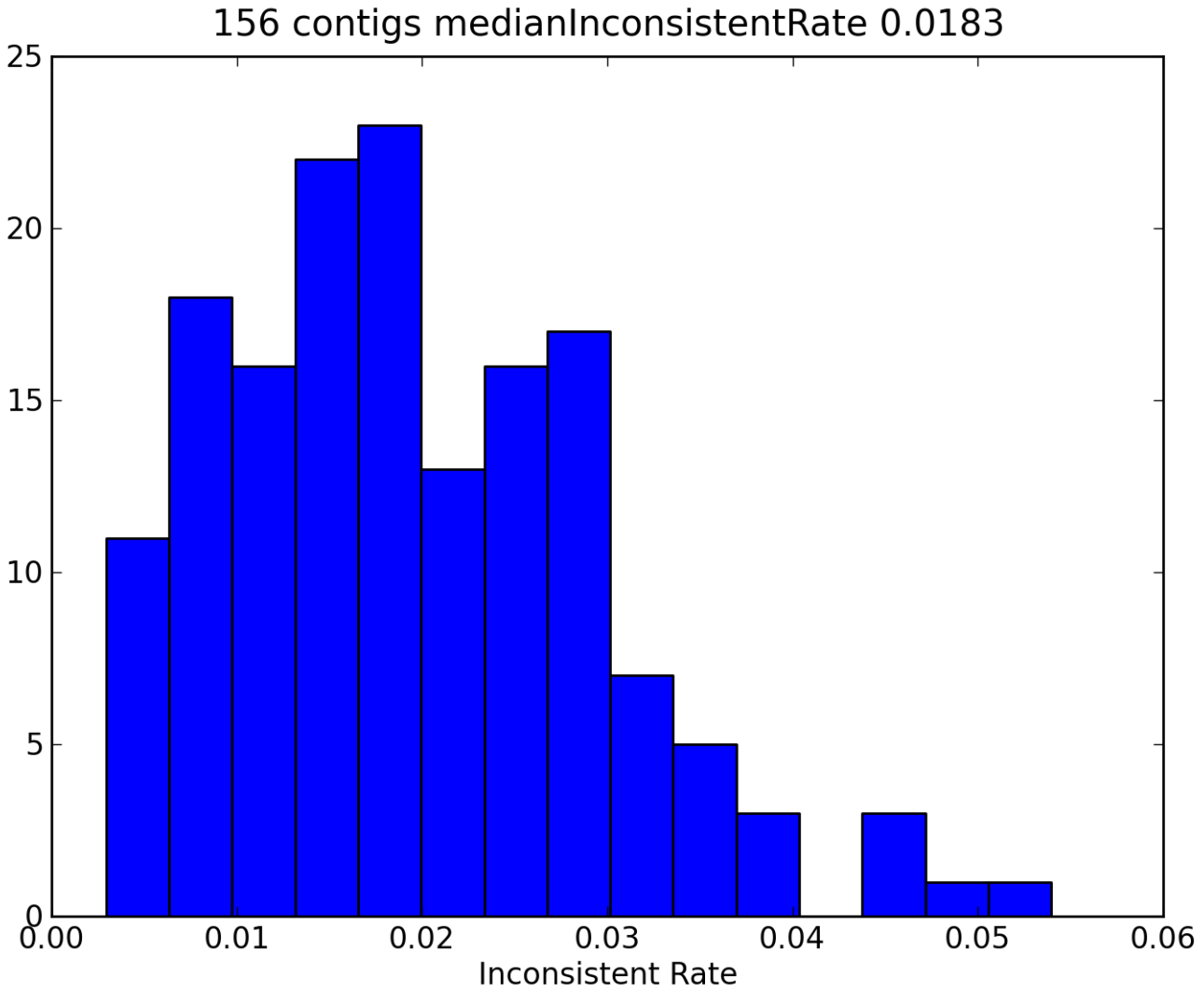
- 4 trios out of 116 VRC sequenced. Two trios have the same parents. => 10 monkeys in total.
- For each locus, if the child's genotype is not one of the 4 combinations of two parents' genotype, this locus is flagged as inconsistent.
- Inconsistent rate = (number of inconsistencies)/(number of total variants in that contig)

Histogram of inconsistent rate for one trio.

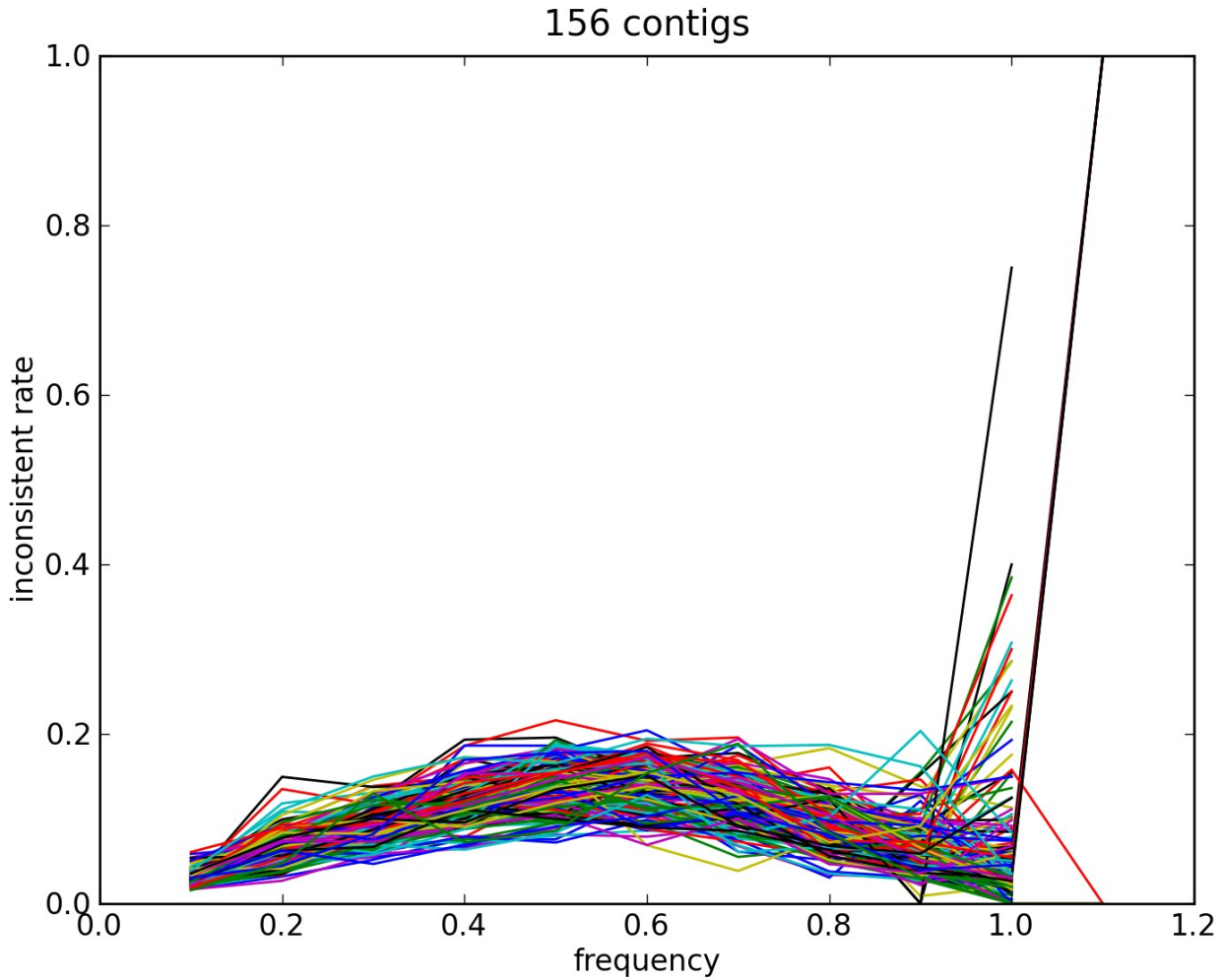


Note: Data includes the **heterozygous** calls.
(For reference, the 1000-genome project gets ~5%-30% error rate in heterozygous calls depending on allele frequency.)

If we focus on homozygous calls only,



Inconsistent rate depends on the AAF (Alternative Allele Frequency)



The sudden increase in the variation of inconsistent rate near 1.0 is due to insufficient number of sites.
The point beyond 1.0 is due to float precision issue (i.e. 1.00001 is regarded as above 1.0).

Variant Density in different populations

Monkey Population	Number of variants per kb
15 Distant VRC monkeys	4.5
5 St. Kitts	3.4
5 Nevis	3.2
5 African subspecies	8.1

Relationship of Carribean monkeys to African subspecies

- Modify the genotype caller to output all sites (both polymorphic and non-polymorphic).
- All heterozygous sites are masked as missing.
- For any pair of two individuals, calculate the distance = (number of sites with different calls)/ (number of total non-NA sites).
- The distance is also calculated between the reference and an individual.

One example: Contig 9

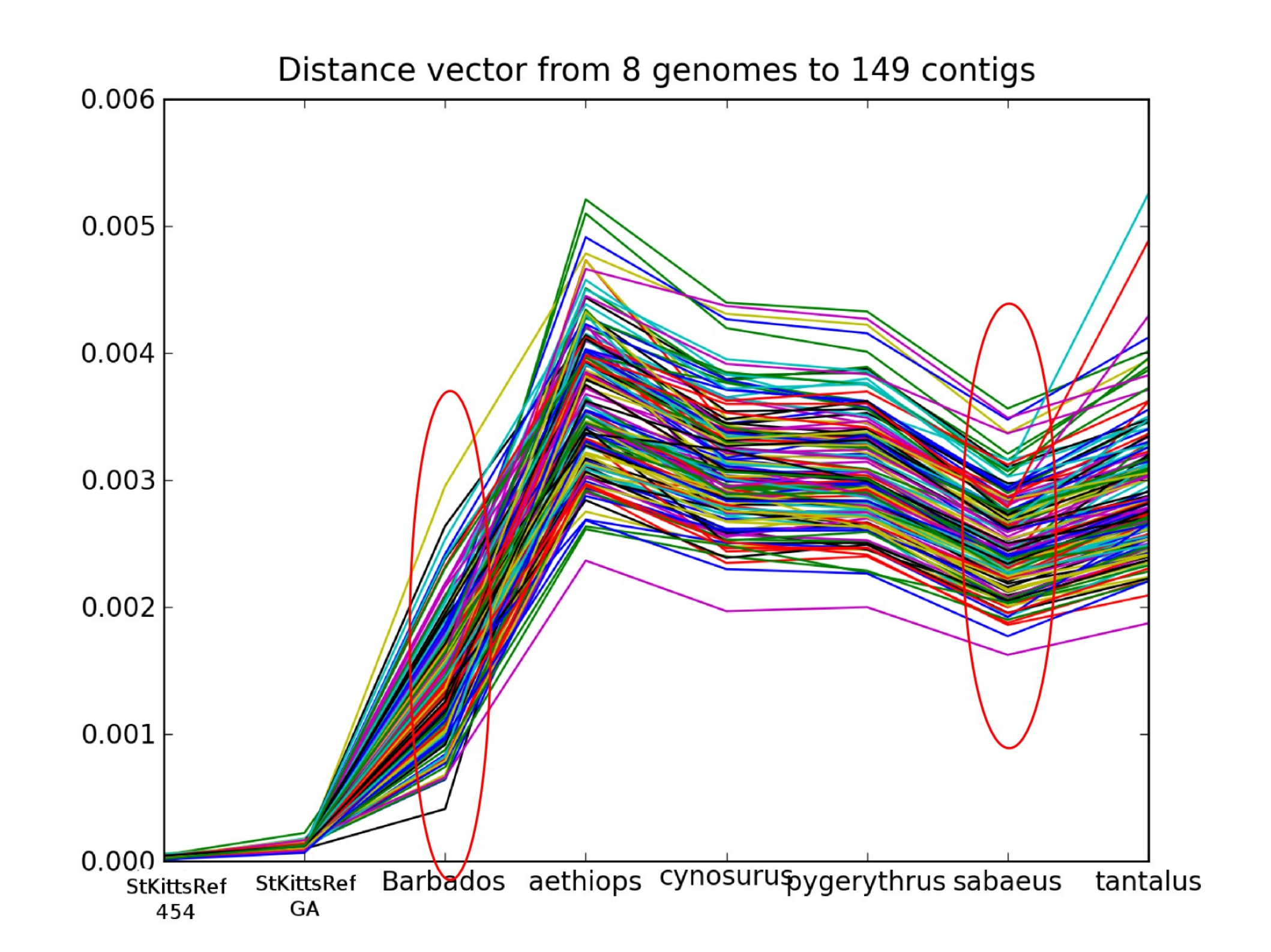
	Contig9	Barbados	StKitts_ref_454	StKitts_ref_GA	aethiops	cynosurus	pygerythrus	sabaeus	tantalus
Contig9		1.64E-03	1.75E-05	1.21E-04	3.45E-03	3.06E-03	3.01E-03	2.38E-03	2.70E-03
Barbados	1.64E-03		1.42E-03	1.39E-03	2.94E-03	2.57E-03	2.52E-03	1.92E-03	2.19E-03
StKitts_ref_454	1.75E-05	1.42E-03		5.17E-05	3.19E-03	2.81E-03	2.75E-03	2.14E-03	2.44E-03
StKitts_ref_GA	1.21E-04	1.39E-03	5.17E-05		3.19E-03	2.81E-03	2.75E-03	2.14E-03	2.44E-03
aethiops	3.45E-03	2.94E-03	3.19E-03	3.19E-03		1.70E-03	1.59E-03	1.28E-03	1.71E-03
cynosurus	3.06E-03	2.57E-03	2.81E-03	2.81E-03	1.70E-03		5.80E-04	5.37E-04	1.11E-03
pygerythrus	3.01E-03	2.52E-03	2.75E-03	2.75E-03	1.59E-03	5.80E-04		3.72E-04	1.05E-03
sabaeus	2.38E-03	1.92E-03	2.14E-03	2.14E-03	1.28E-03	5.37E-04	3.72E-04		4.66E-04
tantalus	2.70E-03	2.19E-03	2.44E-03	2.44E-03	1.71E-03	1.11E-03	1.05E-03	4.66E-04	

Red: error rate.
Yellow: closest pair.
Cyan: 2nd-closest pair.

Draw the distance-vector to reference

	Contig9	Barbados	StKitts_ref_454	StKitts_ref_GA	aethiops	cynosurus	pygerythrus	sabaeus	tantalus
Contig9		1.64E-03	1.75E-05	1.21E-04	3.45E-03	3.06E-03	3.01E-03	2.38E-03	2.70E-03

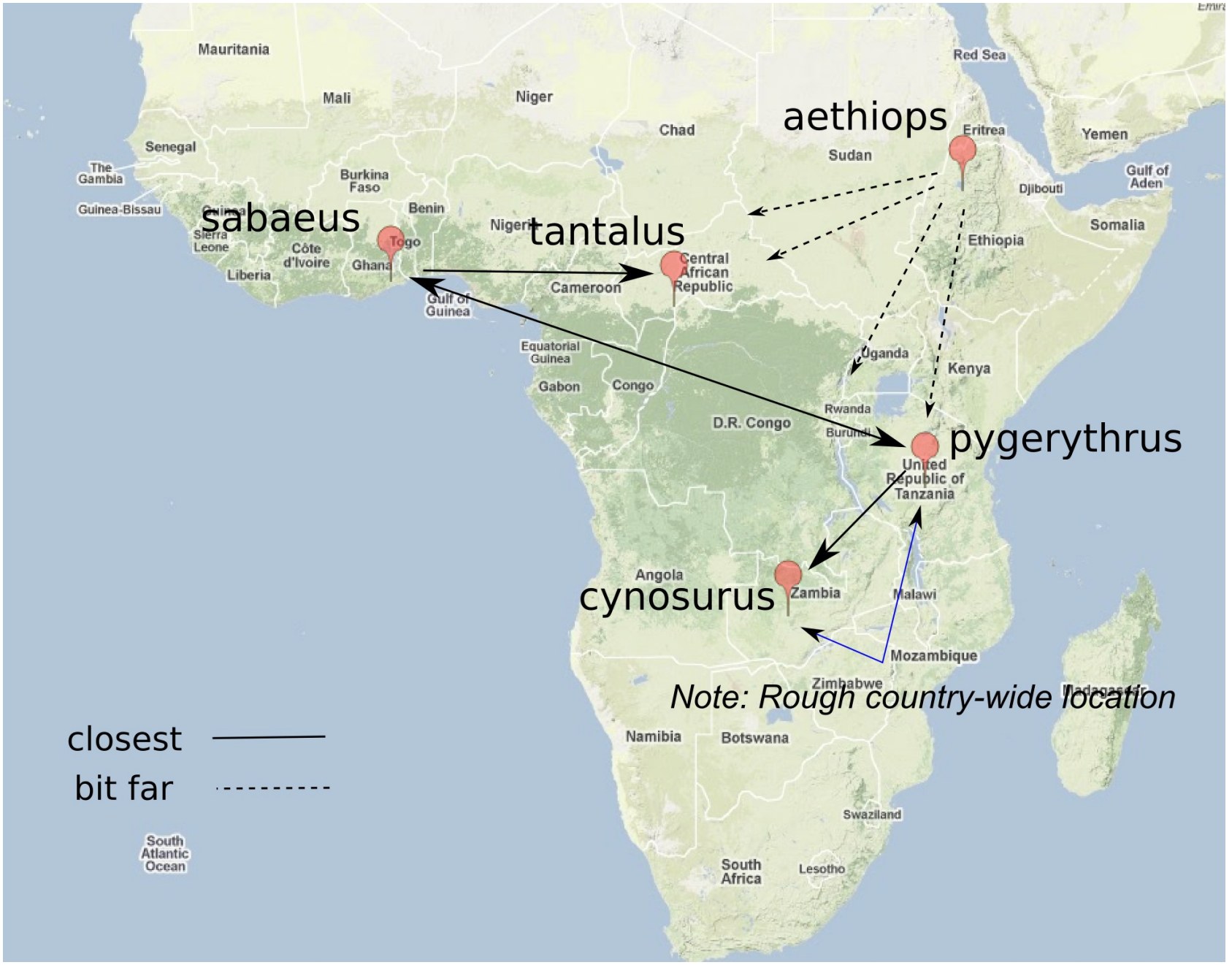
Draw vectors from all contigs in one plot.


Distance vector from 8 genomes to 149 contigs
0.006
0.005
0.004
0.003
0.002
0.001
0.000
StKittsRef 454
StKittsRef GA
Barbados
aethiops
cynosurus
pygerythrus
sabaeus
tantalus

Relationship of Carribean monkeys to African subspecies

- Barbados is much closer to the St. Kitts reference than other African subspecies.
- Sabaeus is consistently the closest African subspecies to St. Kitts reference.
- What is causing that big variation among different contigs?
- Why the order based on distance to ref is changed in some contigs?

Relationship among the African subspecies



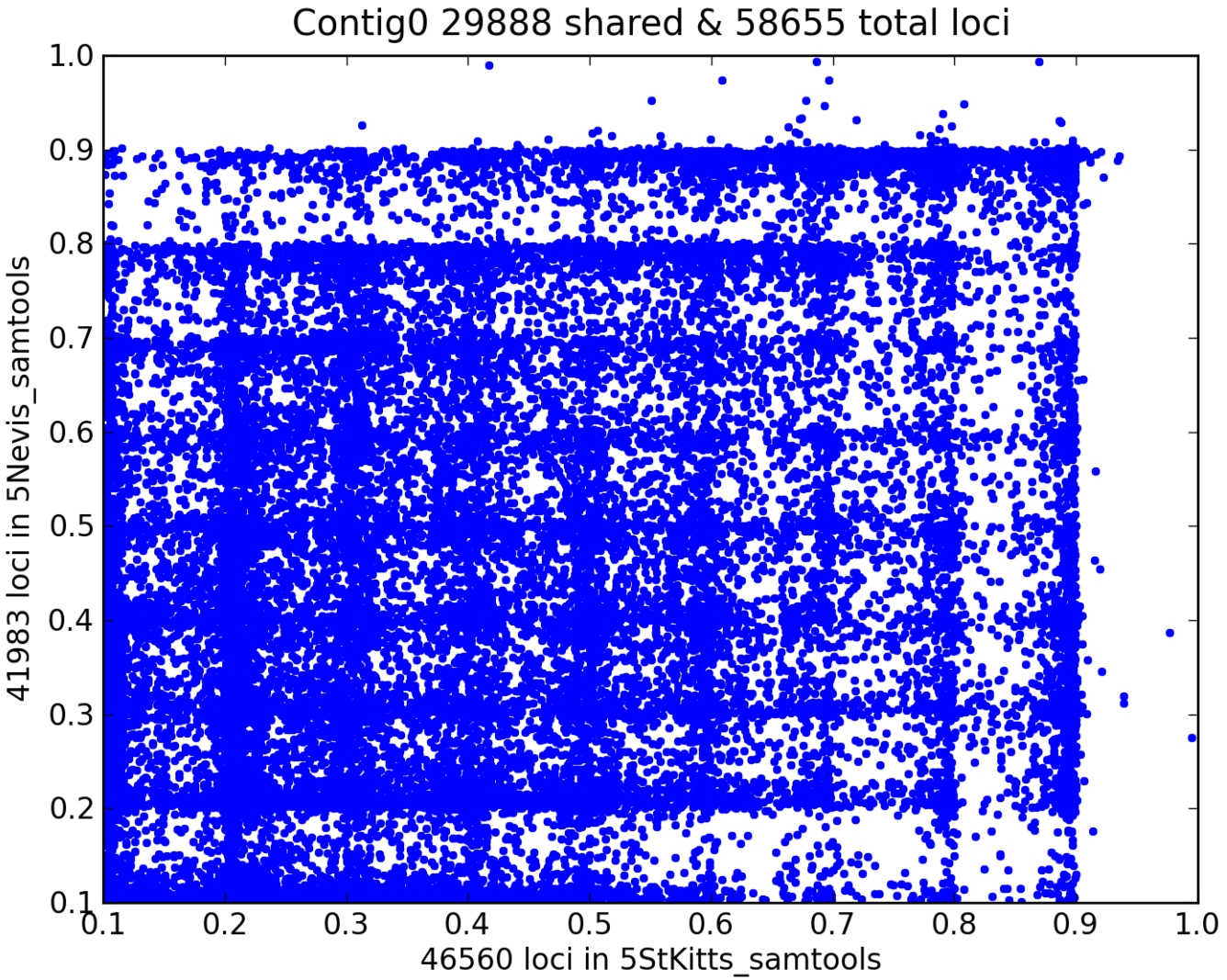
Relationship among the African subspecies

- For pygerythrus, sabaeus has consistently been its closest subspecies.
- However, cynosurus is geographically its closest, at least based on the current rough location.

St. Kitts v.s. Nevis

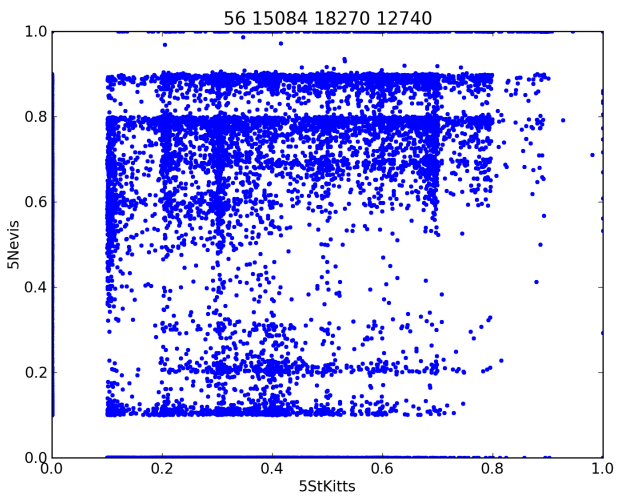
- 5 St. Kitts and 5 Nevis wild monkeys have been sequenced.
- We are wondering how close they are to each other. Is there still genetic exchanges between them?
- We ran the variant-calling pipeline on the two populations separately and compare the AAF (Alternative Allele Frequency).

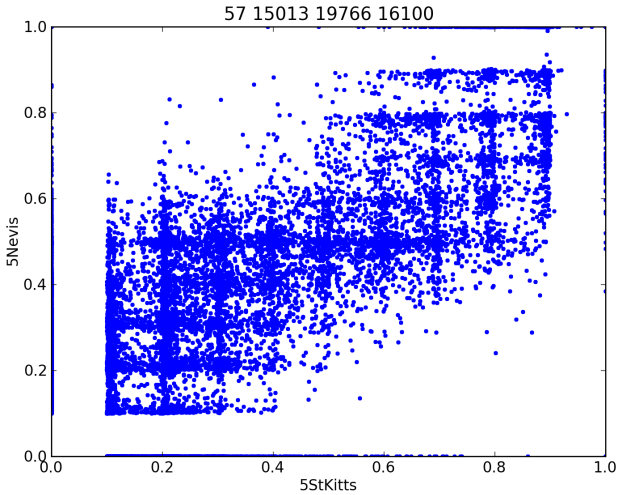
Most contigs look like this:

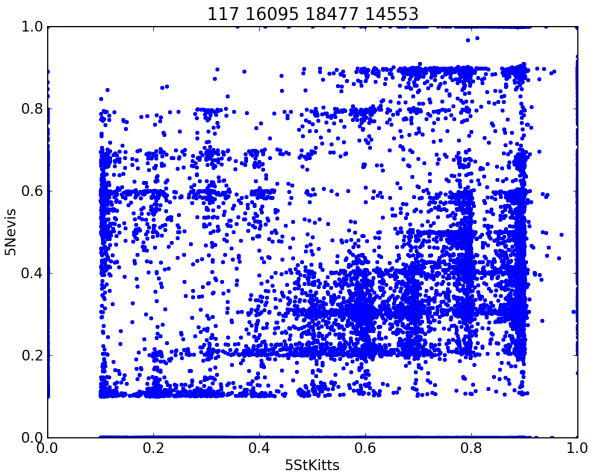


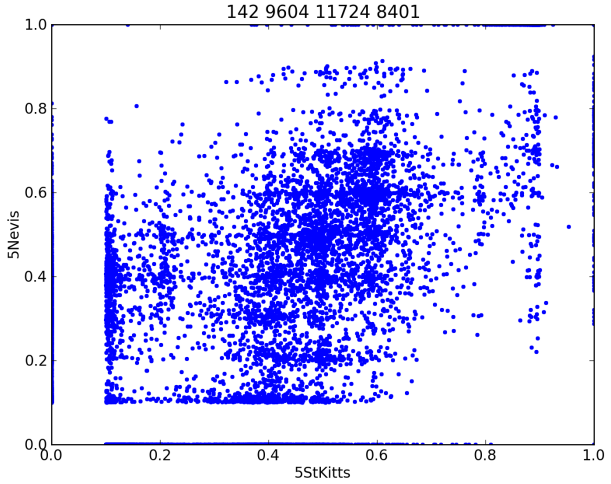
Each dot is a SNP locus. Plotted are the 29,888 shared polymorphic variants.
16,672 (~35.8%) St. Kitts variants are not polymorphic in 5 Nevis monkeys.
The number for the **Nevis** variants is **12,095 (28.8%)**.

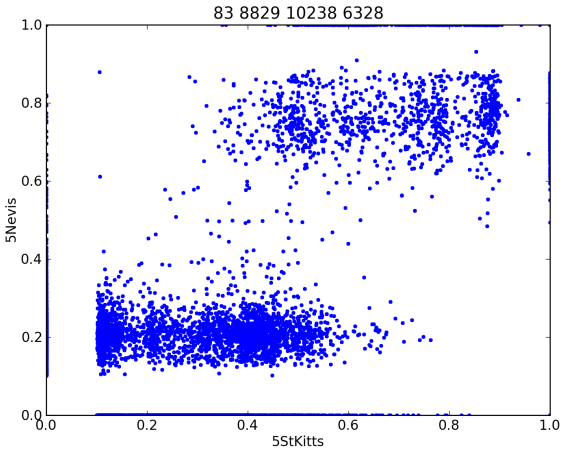
But some contigs look like:

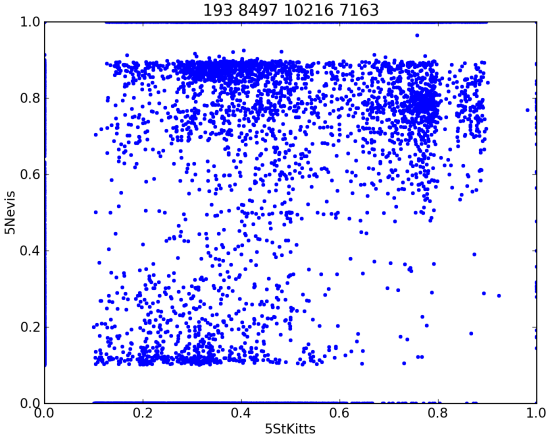






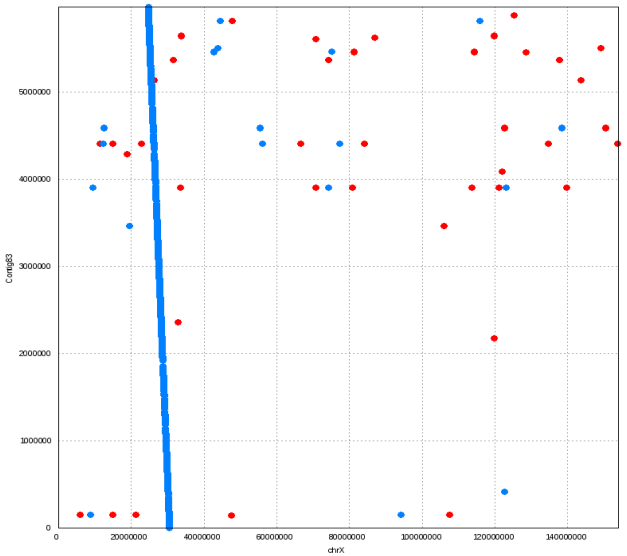




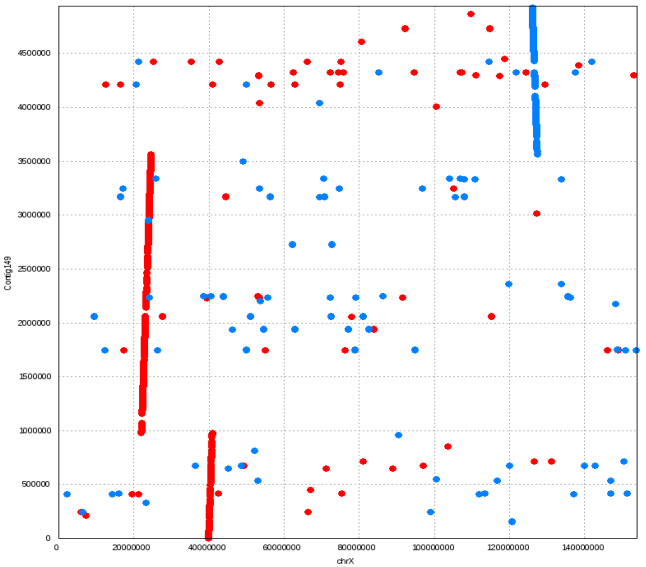


Focus on three contigs that are mapped to chr X.

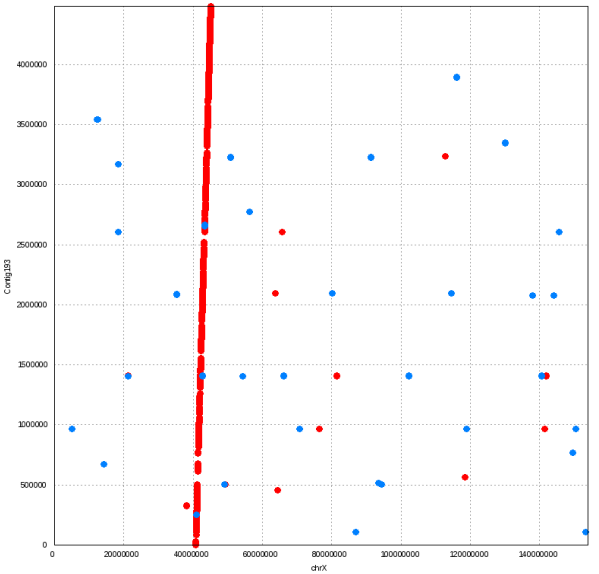
Contig 83



Contig 149

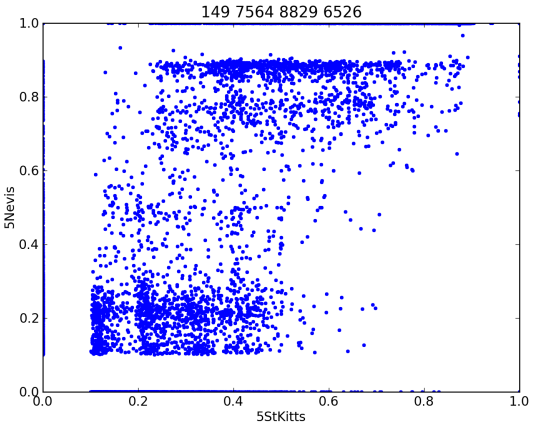


Contig 193



Macaque ChrX

Their AAF plots all look like



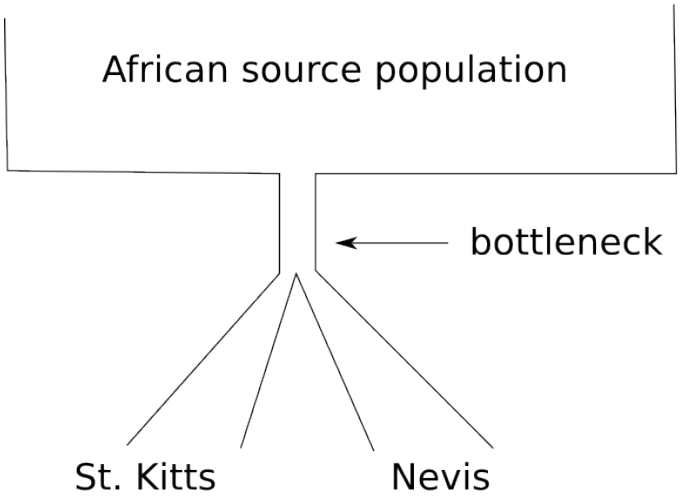
Chr X in 5 Nevis monkeys is of 2 haplotypes.

What are they suggesting?

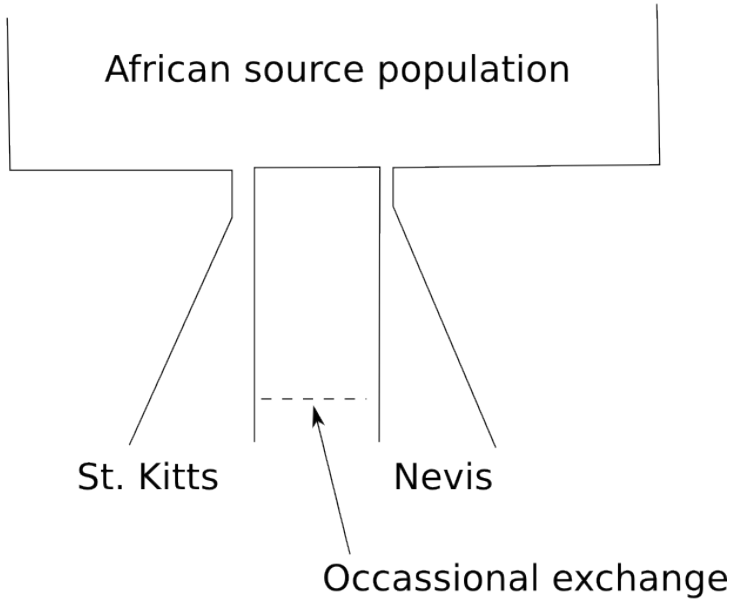
- St. Kitts and Nevis populations are separate and quite different from each other.
- In some genomic regions, there is correlation of AAF, suggesting a common ancestry.
- In some genomic regions, one population (5 monkeys in total though) is composed of 2 or 3 haplotypes, esp. the contigs on chr X.

Potential models to test

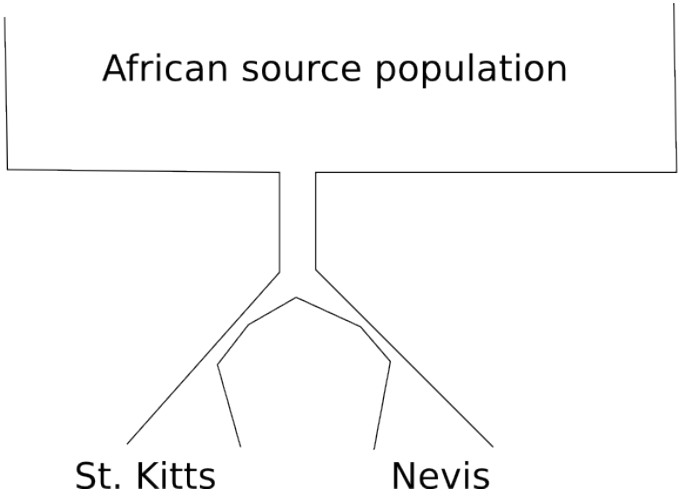
A. Sharing an ancestry



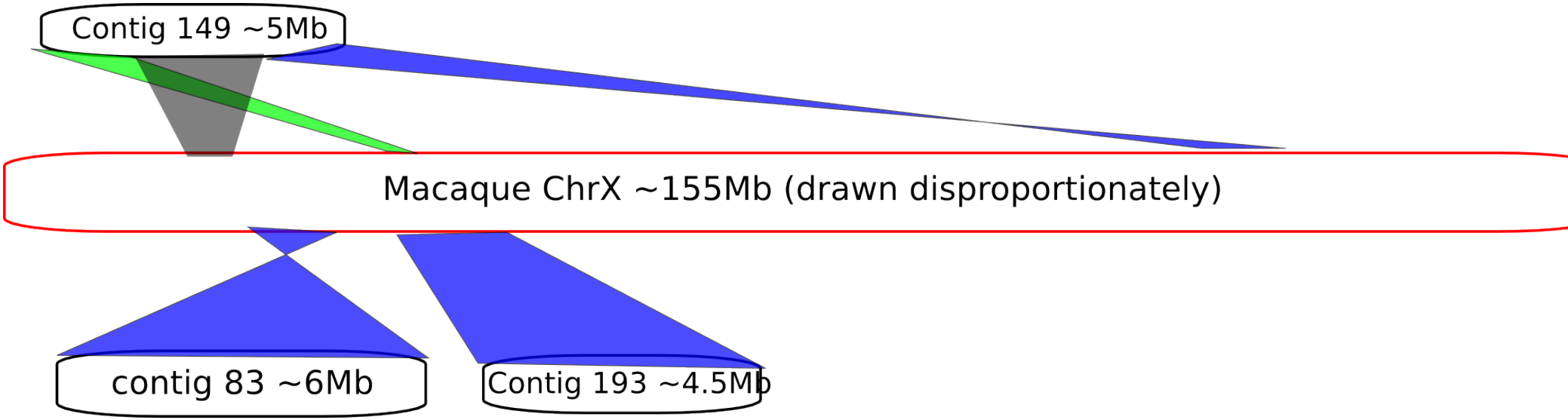
B. No ancestry sharing



C. Sharing an ancestry but each going through additional bottleneck



BTW, the three contigs that are all mapped to chr X suggest this structural event.

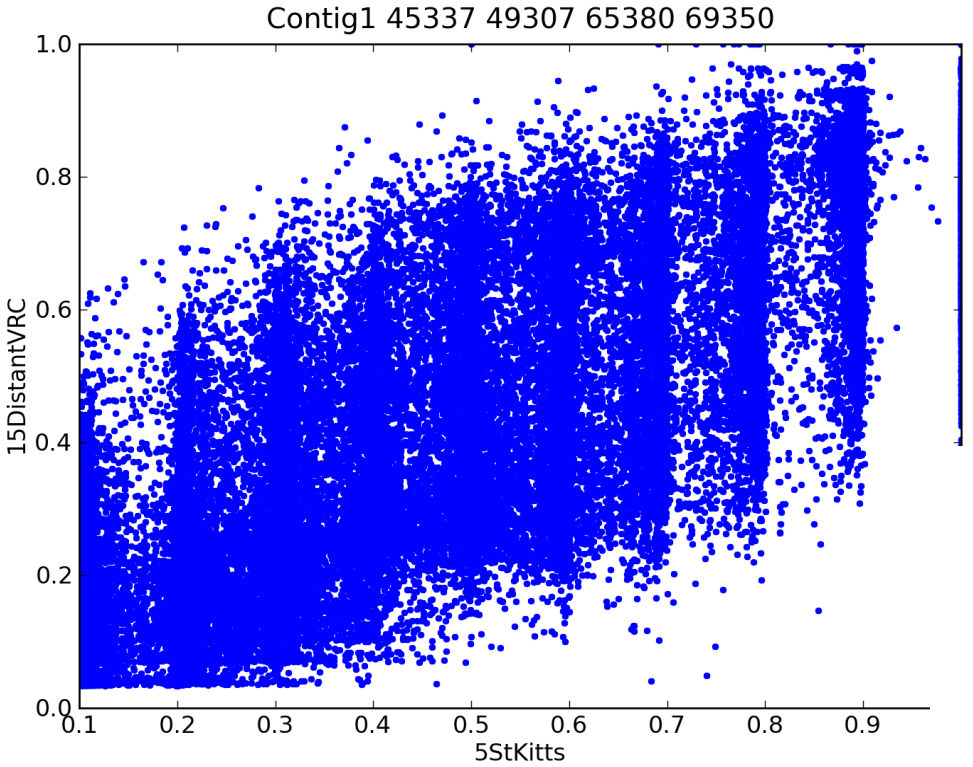


Duplication + inversion + translocation??

VRC v.s. St. Kitts & Nevis

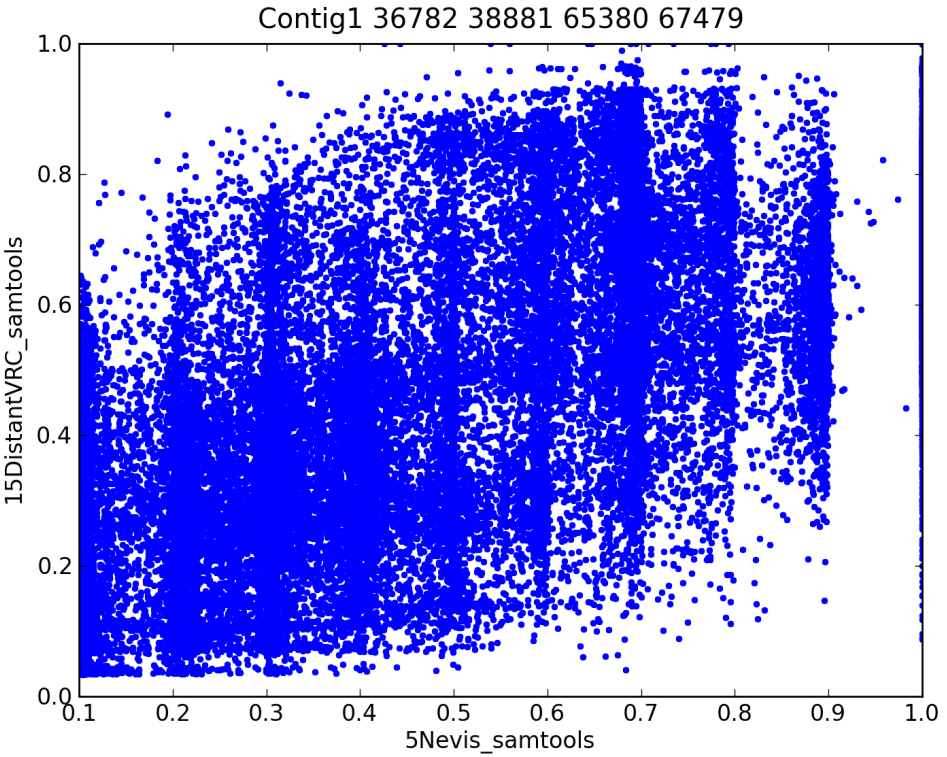
- Mating in VRC is by no means in HWE. We could not use all 116 sequenced VRC monkeys to estimate allele frequency because HWE is assumed in samtools.
- So we picked 15 most-distant ones among the 116 sequenced to form a VRC sub-population.
- We ran the variant-calling pipeline on this 15-distant-VRC population and compare the AAF (Alternative Allele Frequency) with 5 St. Kitts and Nevis monkeys.

Most contigs look like Contig1:



15DistantVRC v.s. 5 St. Kitts

15DistantVRC v.s. 5 Nevis

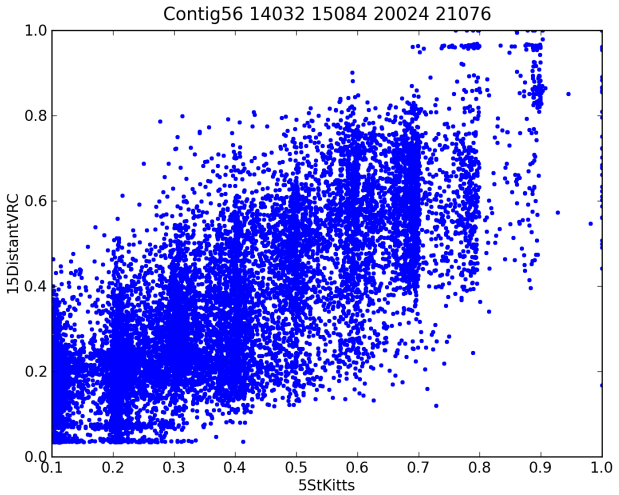


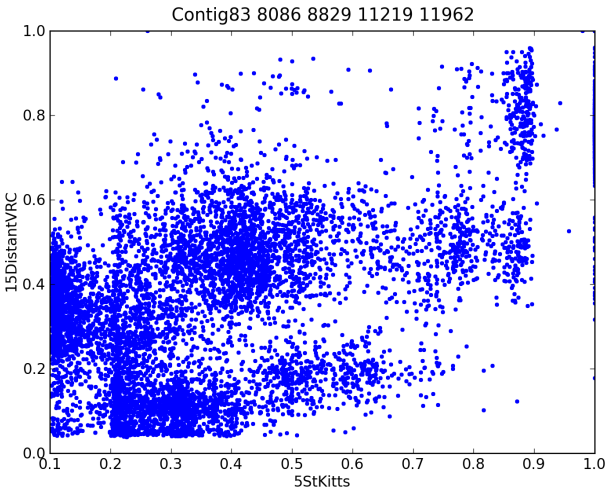
How many variants that are unique to each population for this contig?

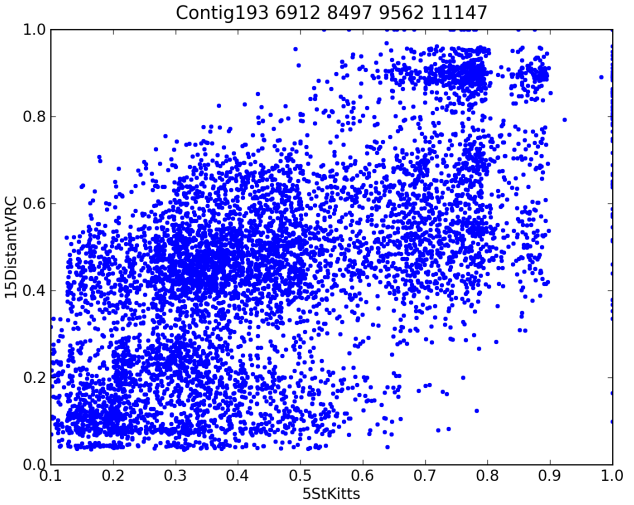
- **5,660 (12.1%) St. Kitts** variants are not polymorphic in 15 distant VRC monkeys. **20,882 (33.8%) for variants from 15 distant VRC** monkeys are not polymorphic in 5 St. Kitts monkeys.
- **4,172 (9.9%) Nevis variants** are not polymorphic in 15 distant VRC monkeys. **23,971 (38.8%) for variants from 15 distant VRC** monkeys are not polymorphic in 5 Nevis monkeys.

A couple of outliers:

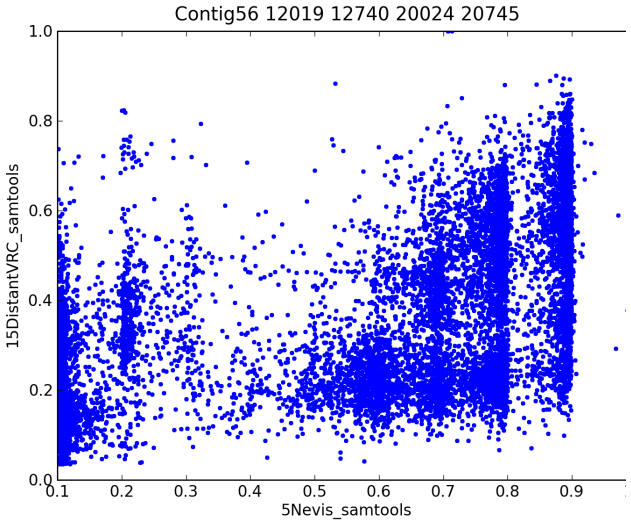
15DistantVRC v.s. 5 St. Kitts

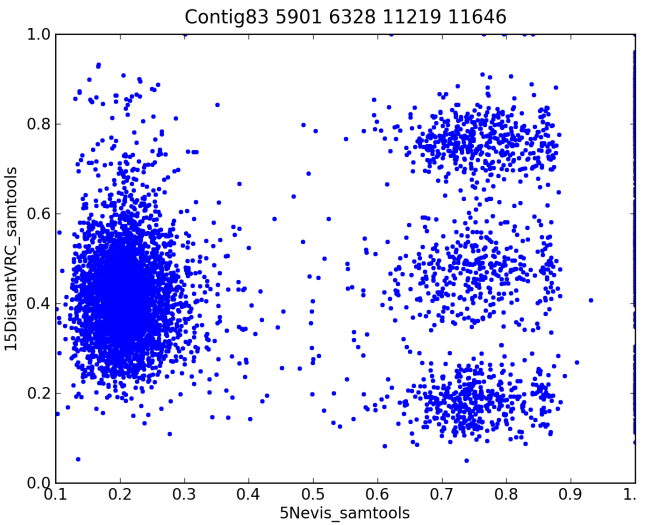


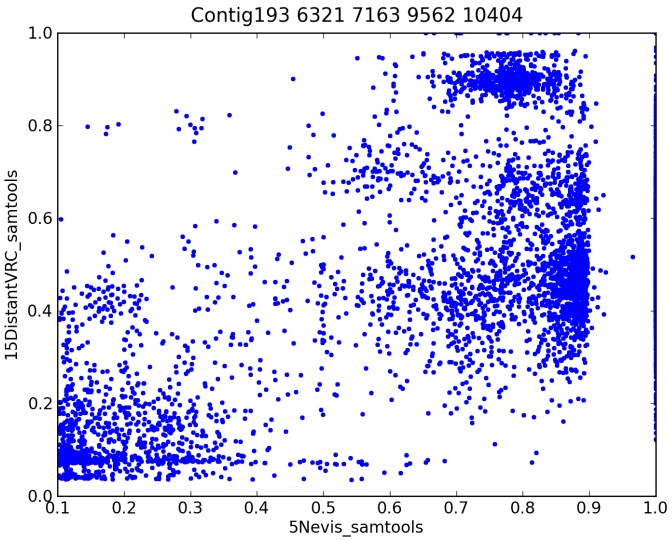




15DistantVRC v.s. 5 Nevis







VRC v.s. St. Kitts & Nevis

- VRC has contribution from both St. Kitts and Nevis population.
- Based on eyeballing AAF correlation, the contribution from St. Kitts is larger than that from Nevis.
- There are regions (esp. chr X) that little correlation exists between VRC and St. Kitts/Nevis.

Outlook

- Refine the pipeline
  - Filter based on depth, strand bias, etc.
  - Try ungapped alignment, which is good for avoiding highly-polymorphic regions.
- Population genetic measures:
  - Nucleotide diversity
  - LD in different populations
  - Fst
  - Estimate ancestral population size, divergence time using human & macaque as outgroup.

Acknowledgement

Everyone in the room, etc.