STANLEY CENTER
FOR PSYCHIATRIC RESEARCH
AT BROAD INSTITUTE

BROAD INSTITUTE

Analyzing genetic diversity in marmoset colonies

Ricardo del Rosario
Computational Biologist
Stanley Center for Psychiatric Research at Broad Institute

- Marmosets are used as a non-human primate model to study autoimmune and infectious disease
- At the Stanley Center for Psychiatric Research at Broad Institute and other research institutes, the marmoset is now being used as a model for psychiatric disorders

- As the demand for marmosets increase, it becomes more important to maintain the genetic diversity of the marmoset colonies in the US and around the world
  - we can use either a genotyping chip or whole genome sequencing

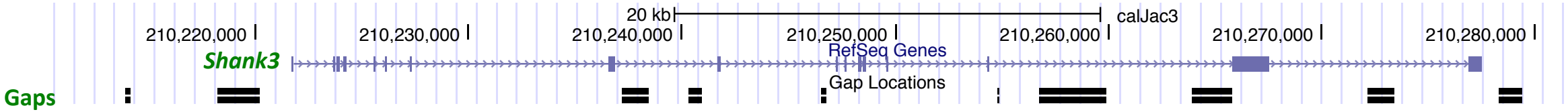
- We have performed whole genome sequencing of 80 marmosets at Broad/MIT
  - the SNP calls can be used a a starting point for designing a genotyping chip
  - we will show how the SNP calls can be used for genetic diversity analysis

- We also present a new marmoset genome assembly that can be used for genomic analyses

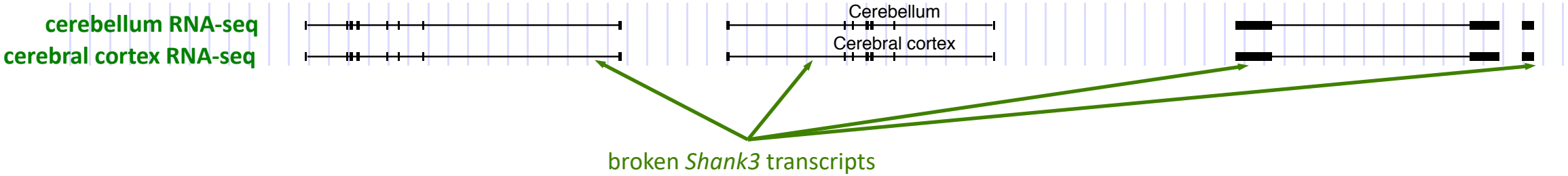
A high-quality reference genome assembly is important for a model organism

In the published assembly, 5.5% of the assembly are gap regions, which disrupt almost all large genes

Ten assembly gaps in the *Shank3* locus



***Shank3* transcripts from RNA-seq pipeline broken by gaps**



Gaps are unknown bases denoted as a sequence of Ns. Gap lengths are estimated

De-novo assembly of the marmoset genome

De novo assembly

sequence DNA fragments

…GCCCCTTTCAGCTTAGGCAGCCTGC

GCATTCCGGCAGGTCGTACACGTACGTACC…

..CGTAGGTCACGAGATATGCGTCTGTAGTACGTG

find overlaps between reads

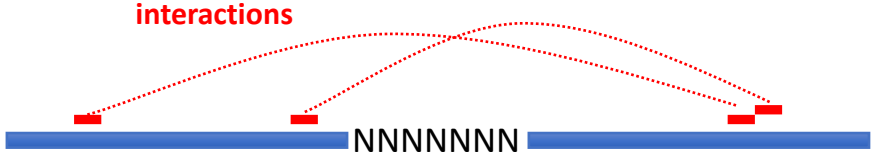
…GCCCCTTTCAGCTTAGGCAGCCTGC

TTTCAGCTTAGGCAGCCTGCACTTGAC

contig

Scaffolding

paired-end reads from long-range DNA interactions



contig gap contig

Illustration of a scaffold

Gap-filling

contig gap contig

CCTTGCATG NNNNNNNN AGGCTTCA

CATG ACCTGCGGT AGGC

contig from another assembly

CCTTGCATG ACCTGCGGTAGGCTTCA

De novo assembly

2x250 PE Illumina library, 60X coverage
Software: Discovar de-novo

Contig N50: 74.4 Kb
Scaffold N50 99.8 Kb

Scaffolding

Chicago library and Hic
Software: HiRise

Contig N50: 81.1 Kb
Scaffold N50: 129.2 Mb

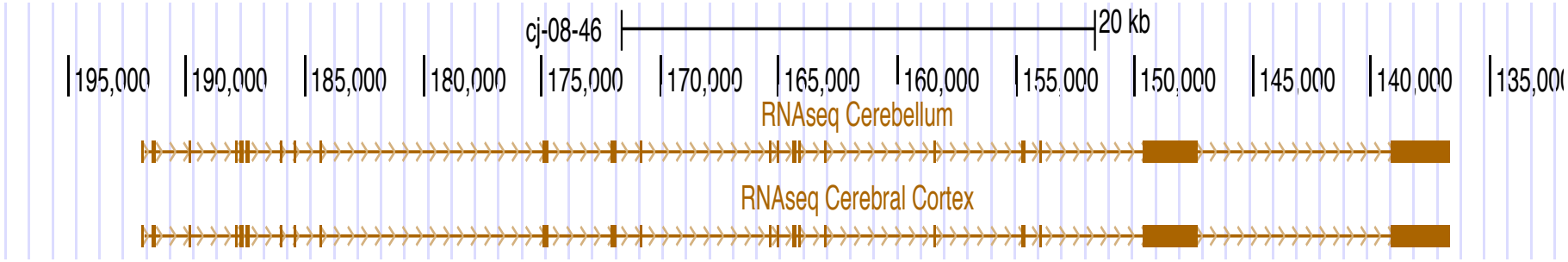
Gap-filling

Map an external contig set to the assembly to fill gaps
Software: GMcloser

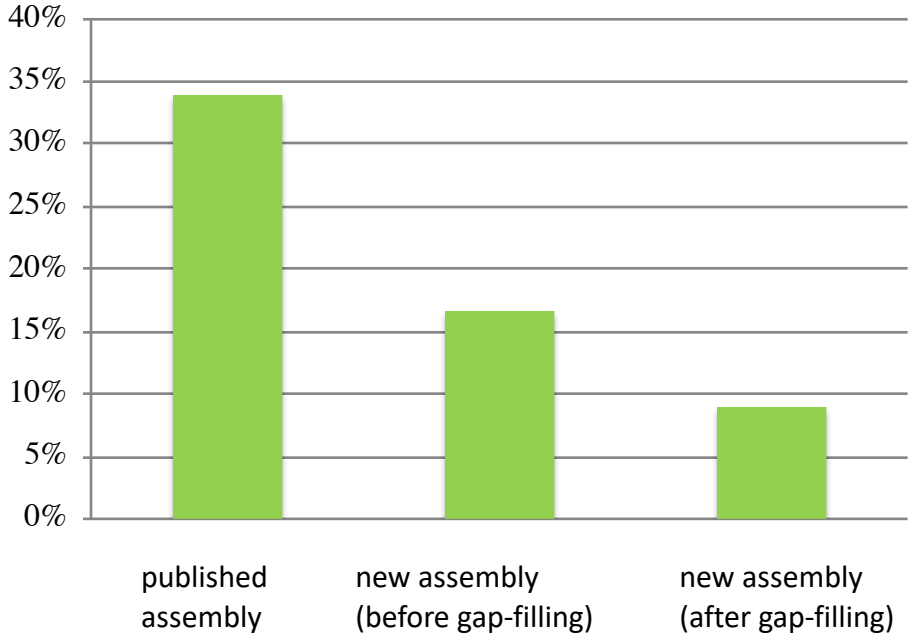
Contig N50: 155.3 Kb
Scaffold N50: 129.2 Mb

The new assembly is 4-fold more contiguous than the reference assembly

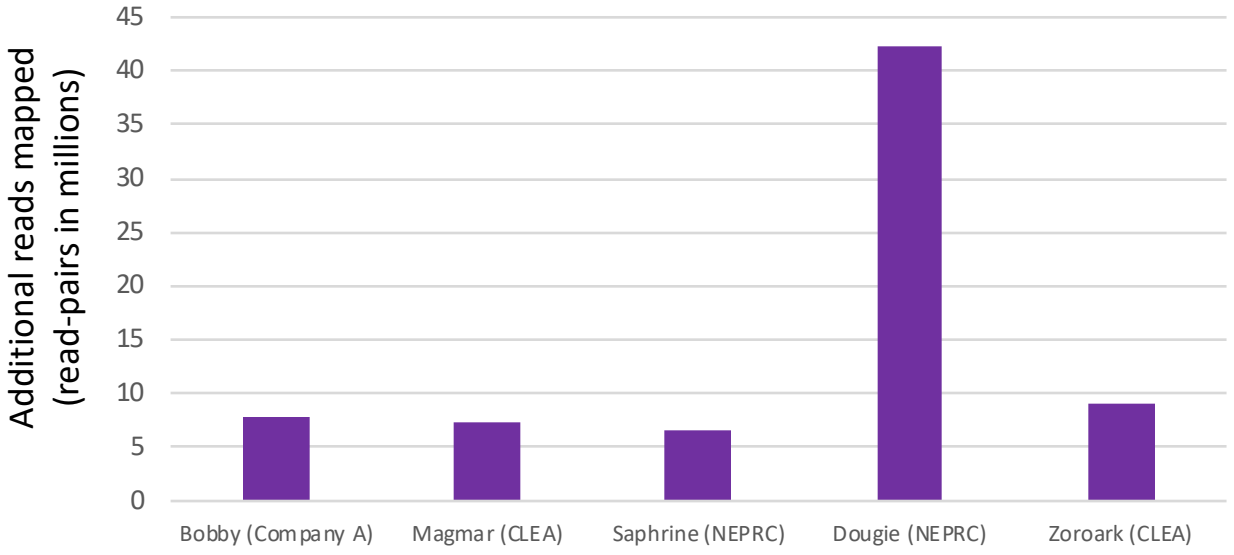
New assembly: *Shank3* RNA-seq transcripts not broken by gaps



Fraction of Transcripts with Gaps



More reads mapped to the new assembly



The new assembly now available at NCBI and annotated by Ensembl

ASM275486v1

Organism name: Callithrix jacchus (white-tufted-ear marmoset)
Infraspecific name: Breed: NEPRC
Isolate: CJ-08-46
Sex: female
BioSample: SAMN07586072
BioProject: PRJNA401030
Submitter: Broad Institute
Date: 2017/11/06
Assembly level: Scaffold
Genome representation: full
GenBank assembly accession: GCA_002754865.1 (latest)
RefSeq assembly accession: n/a
RefSeq assembly and GenBank assembly identical: n/a
WGS Project: NTIC01
Assembly method: Discovar denovo v. 52488; Chicago/HiRise v. 1.3.0-117-g854f52f; HiC/HiRise v. 2.1.2-ad17ecf8bf57
Expected final version: no
Reference guided assembly: de-novo
Genome coverage: 60.0x
Sequencing technology: Illumina HiSeq

See Genome Information for **Callithrix jacchus**

There are 4 assemblies for this organism
See more

IDs: 1432951 [UID] 5620578 [GenBank]

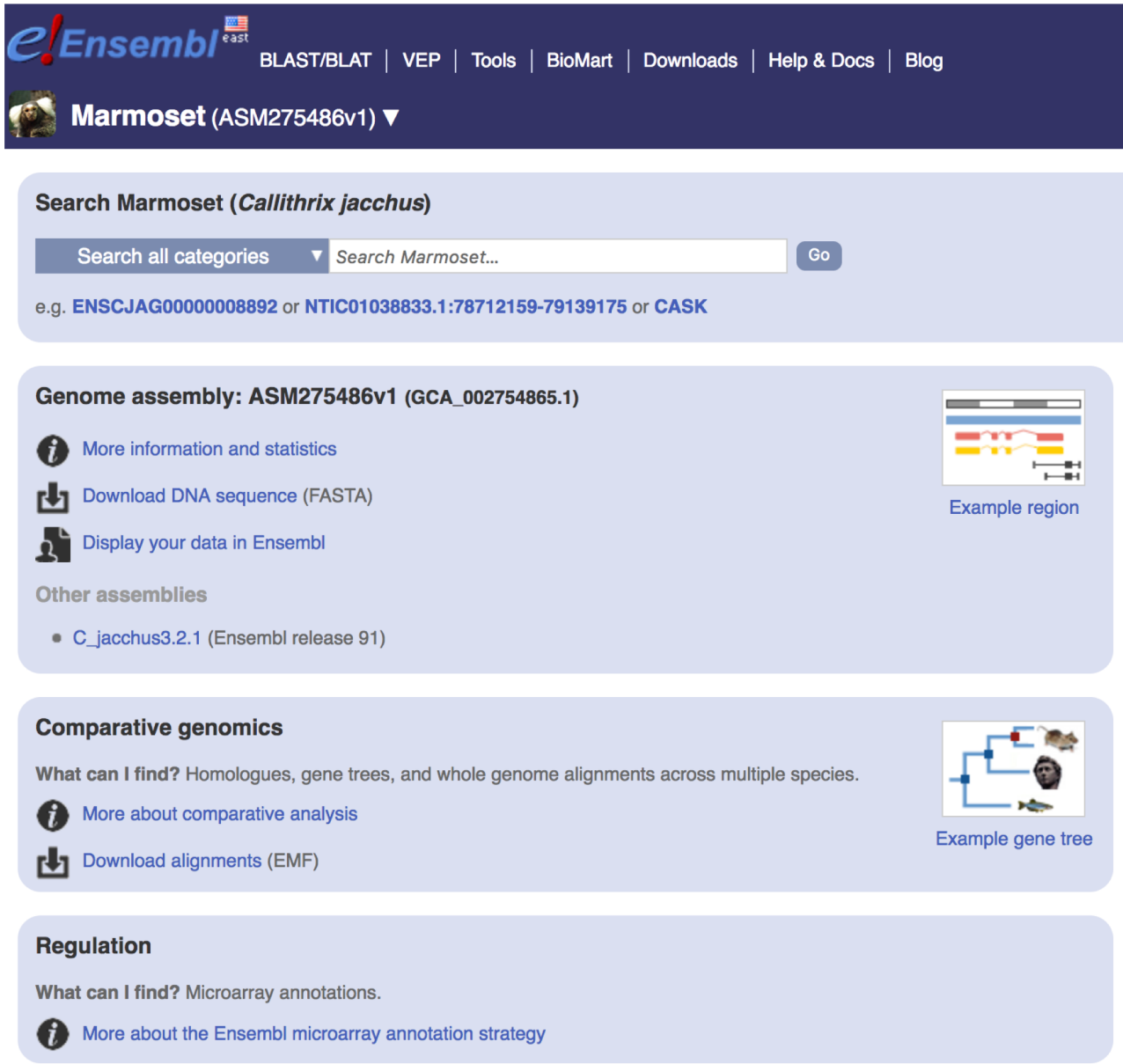
History (Show revision history)

Comment

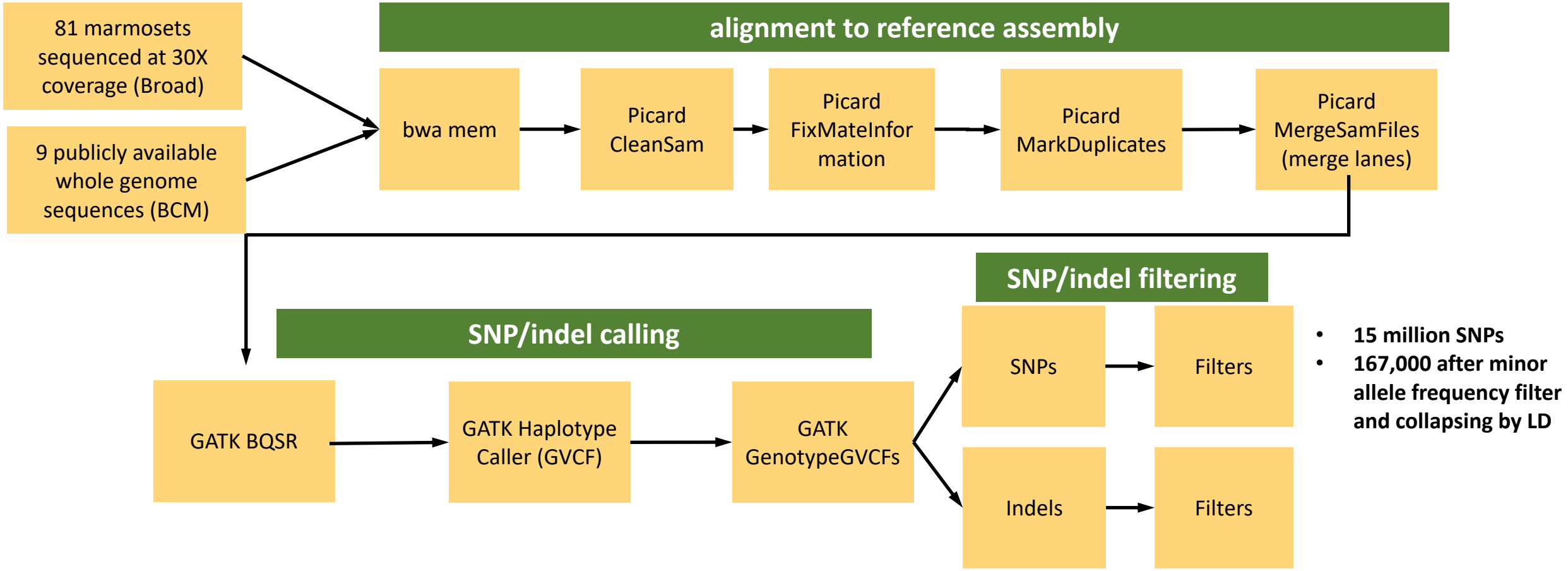
All assembly gaps are from paired-end information, but from different technologies. 100 bp gaps are from scaffolding by Discovar de-novo, 200 bp gaps are from scaffolding by Chicago/HiRise (Dovetail Genomics), and 300 bp gaps are from scaffolding by Hic/HiRise ... more

Global statistics

Total sequence length	2,845,375,248
Total assembly gap length	9,511,100
Gaps between scaffolds	0
Number of scaffolds	39,944
Scaffold N50	129,239,660
Scaffold L50	9
Number of contigs	88,439
Contig N50	155,284
Contig L50	5,014



SNP detection from whole genome sequences



- These 167,000 SNPs can be used as a starting point to design the genotyping chip, and ideally should be augmented by SNPs from other colonies
- What do the genotypes at these SNP locations tell us about genetic diversity in marmosets?

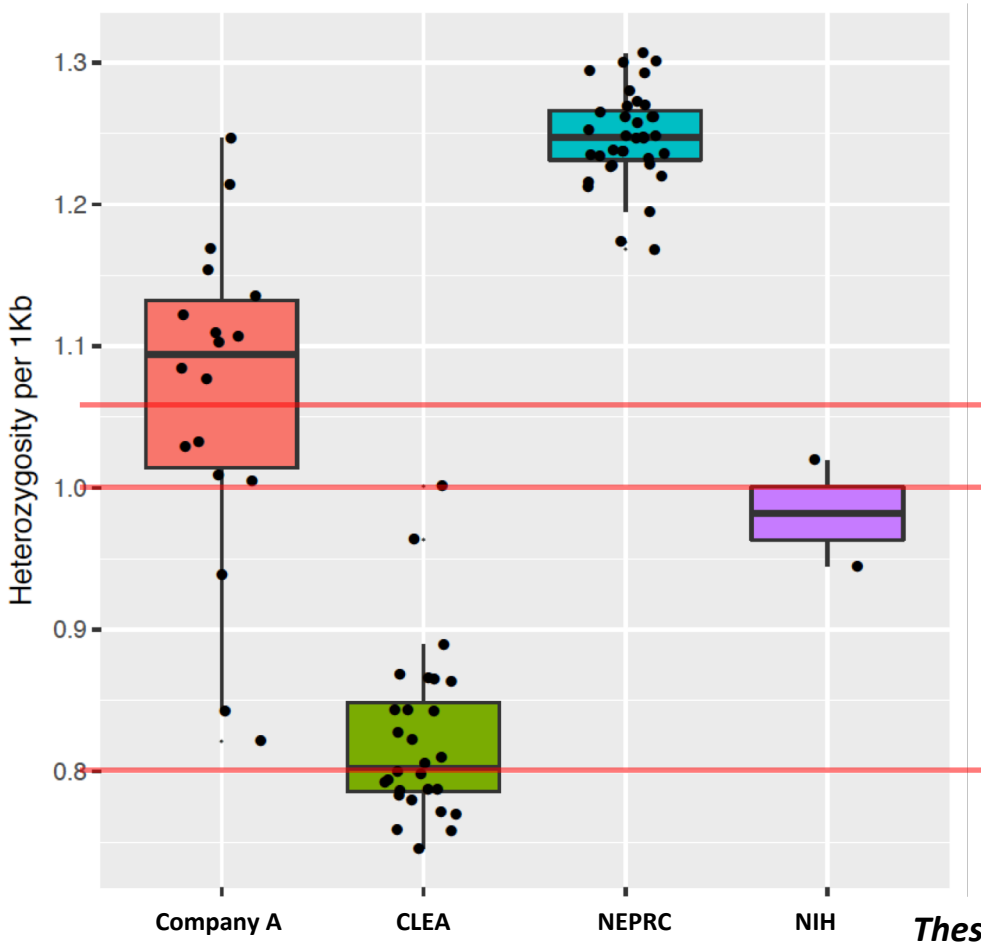
How genetically diverse are the captive marmosets?

Heterozygosity

- a metric for understanding genetic variation in a population
- count nucleotides within an individual that differ between the chromosomes inherited from the parents

from mother ACAGGTACTGACCTACTCCGATCGGACTAGCGATCCTGACTTGCA

from father ACAGGTACGGACCTTCTCCGATAGGACTAGCGATCCTGACTTGCA



ave heterozygosity of all marmosets we sequenced

ave heterozygosity of humans in Africa (1 per Kb)

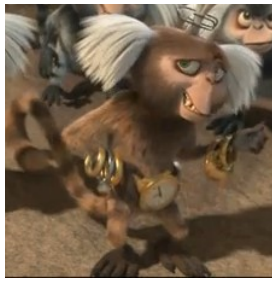
ave heterozygosity of humans in Europe (0.8 per Kb)

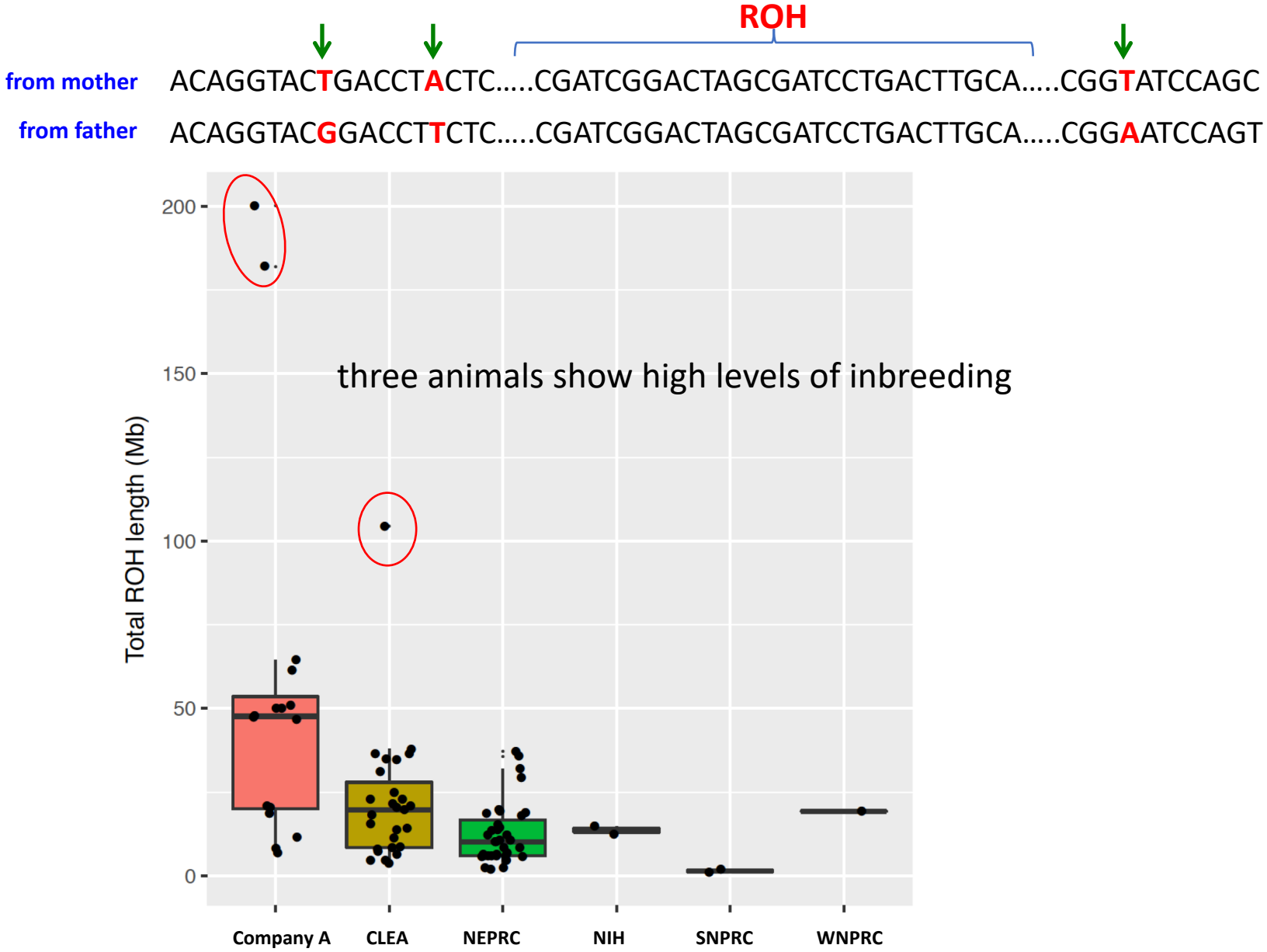
- **NEPRC**: New England Primate Research Center (Broad/MIT)
- **CLEA** (marmosets from Japan)

These diversity estimates will change as we improve our SNP calling accuracy

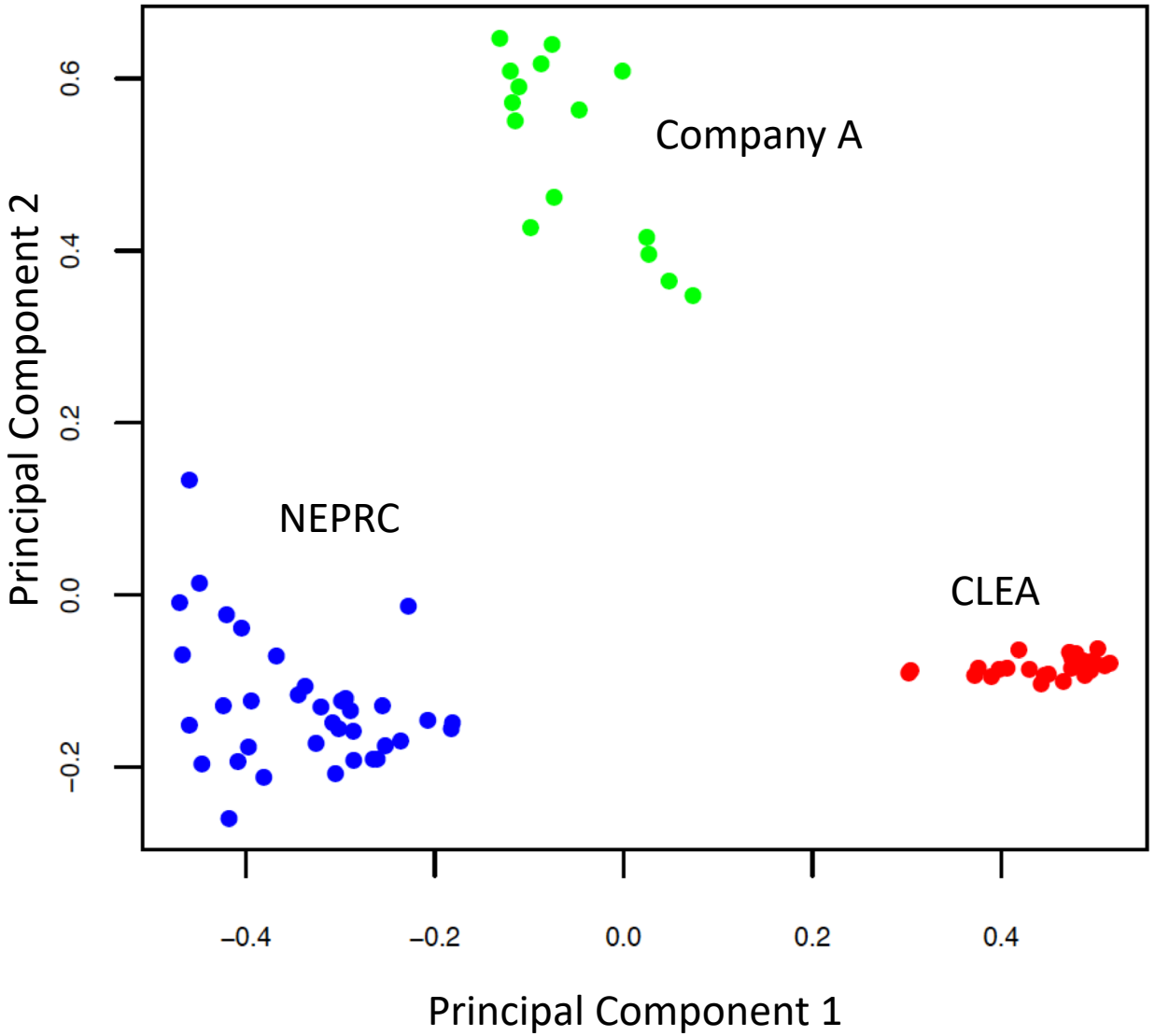
How genetically diverse are the captive marmosets?

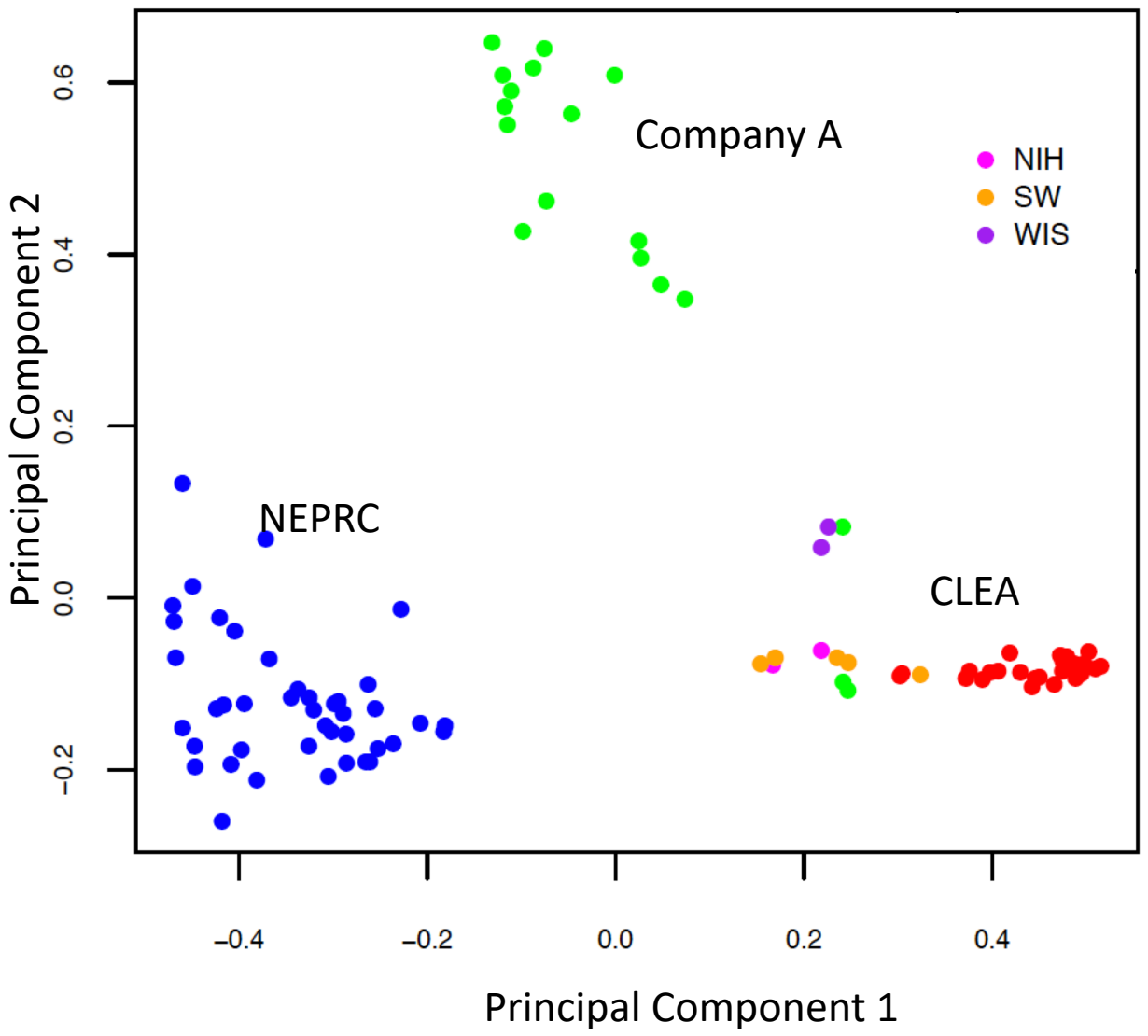
Runs of Homozygosity (ROH): contiguous lengths of homozygous genotypes from identical haplotypes inherited from parents



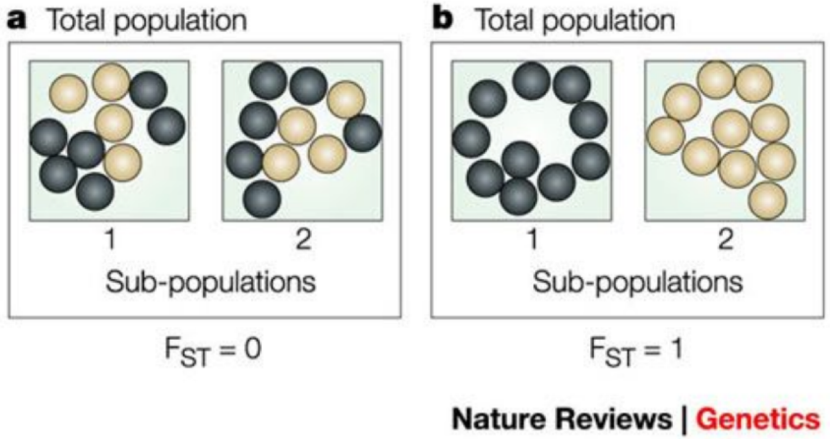


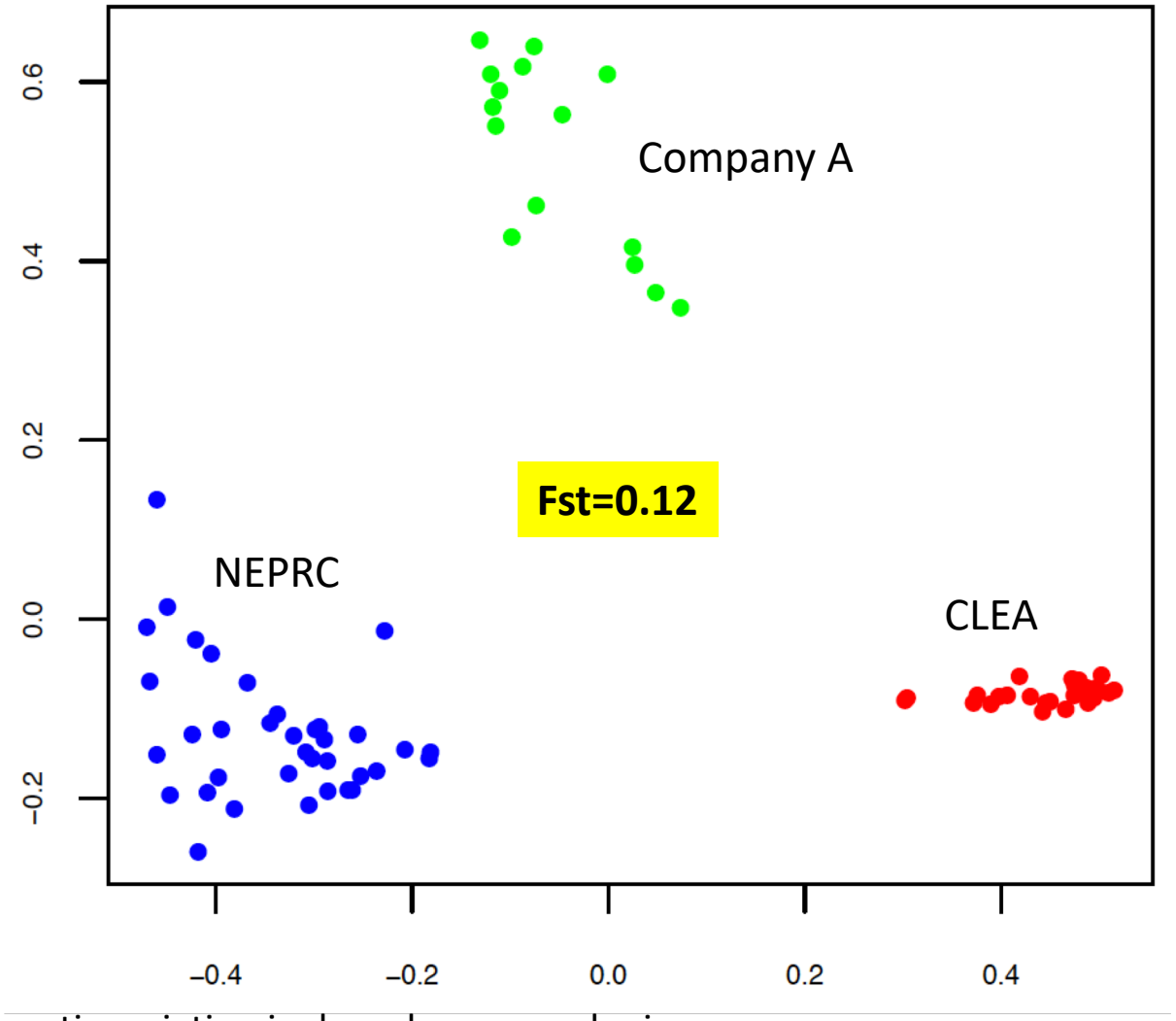
PCA analysis shows that the colonies are genetically distinct





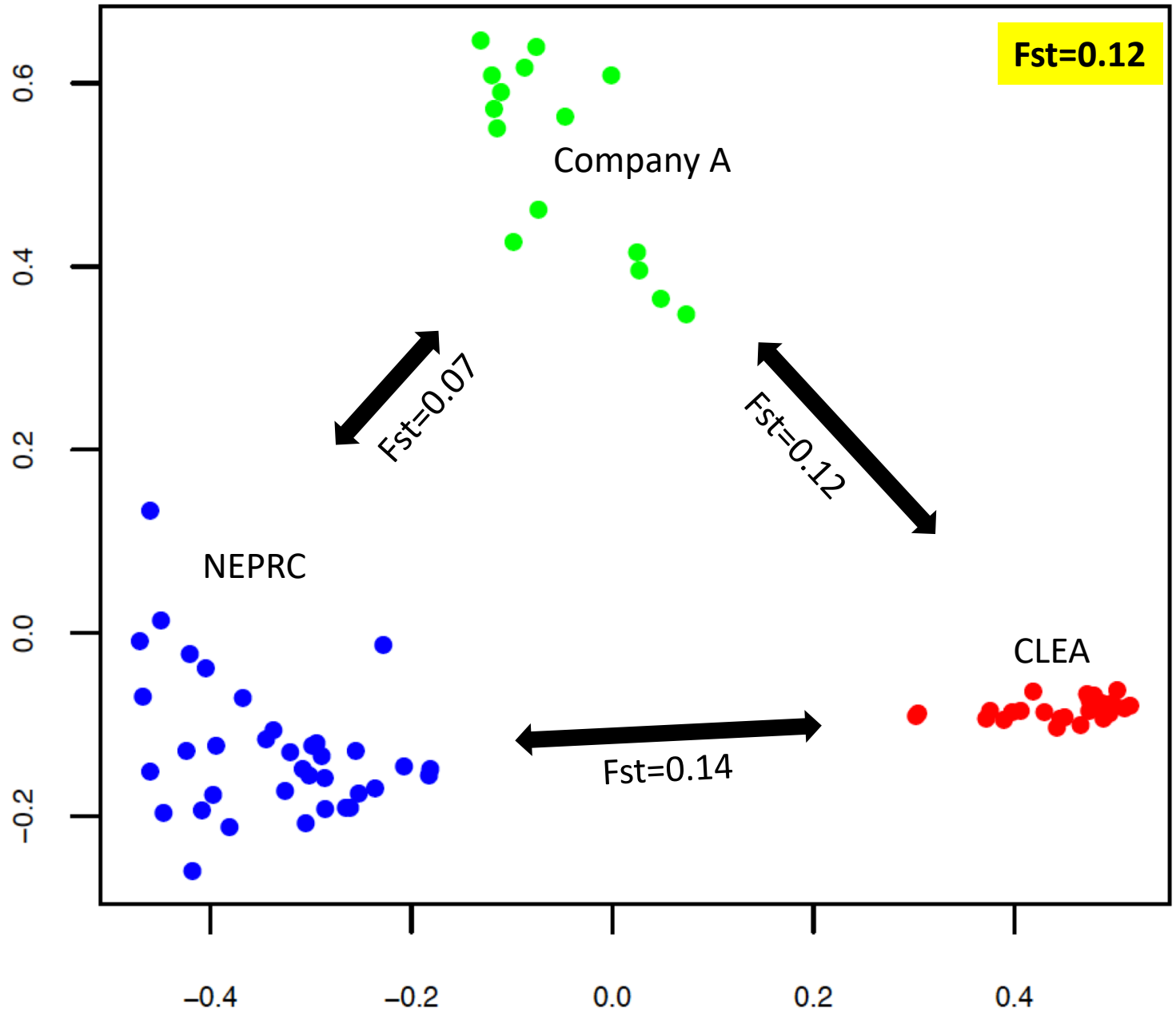
Fst (a measure of structure in populations) of three marmoset colonies



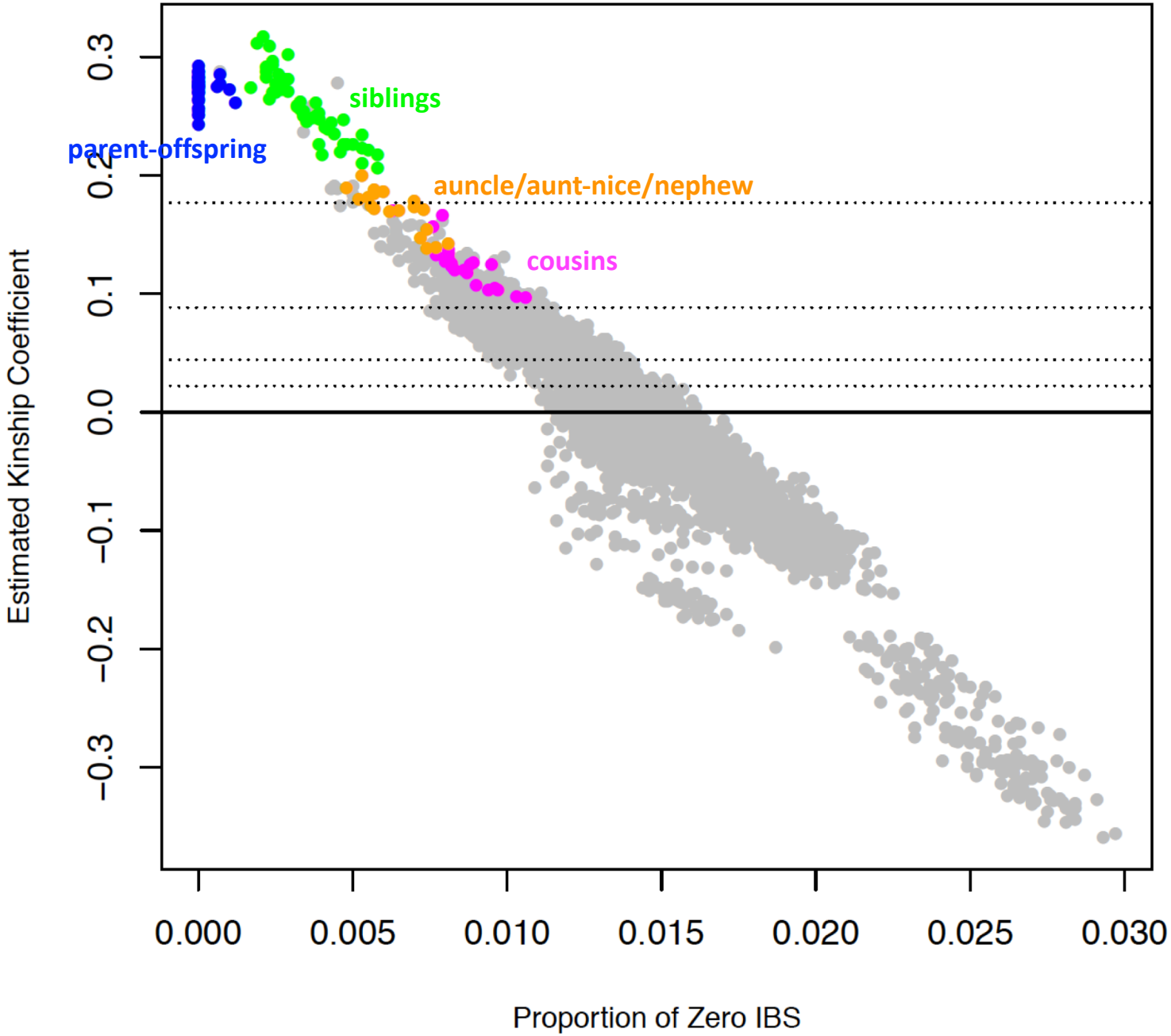


- 88% of the total genetic variation is shared among colonies
- only 12% of the genetic variation come from allele frequency differences between colonies
- In comparison Fst between human populations is 0.15

Fst (a measure of structure in populations) of three marmoset colonies:



SNPs can detect relatedness



coloring:

- known relationships
- not used to calculate kinship coefficients

Summary

- SNPs calculated from whole-genome sequencing allowed us to study genetic variation in marmosets at Broad/MIT (NEPRC, CLEA, Company A)

- Heterozygosity of NEPRC marmosets is higher than in humans; we should keep the level of heterozygosity by careful breeding strategies

- 167,000 SNPs can be used as a starting point for designing a genotyping chip for marmosets
  - include whole genome sequences from marmosets from other colonies
  - genotyping platform: Affymetrix or Illumina?

- Since marmoset blood is chimeric (contains DNA from the twin), which tissues can we use for genotyping and whole genome sequencing? We have been using DNA from cultured fibroblasts to address the contamination issue.

STANLEY CENTER
FOR PSYCHIATRIC RESEARCH
AT BROAD INSTITUTE

BROAD INSTITUTE

Acknowledgements

Steven McCarroll (Harvard Medical School and Broad)
Curtis Melo
Avery Davis Bell
Fenna Krienen
Diane Gage
Anna Neumann
Giulio Genovese
Alec Wysoker

Jessica Alfoldi (Broad)

Broad Genomics Platform

Dovetail Genomics

Guoping Feng (MIT and Broad)
Qiangge Zhang
Yang Zhao
Xian Gao

James Fox (MIT)
Monika Burns
Stephen Artim
Alex Sheh

James Pickell (NIH)