Opportunities to improve genomic information and resources for marmosets

Jeffrey Rogers, Ph.D.
Human Genome Sequencing Center
Baylor College of Medicine
Houston, Texas

Fundamental Resources for Primate Genomics

- Whole Genome Reference Assembly
  - Essentially complete (few gaps) with minimal sequence errors
- Accurate and “Complete” Annotation
  - Protein coding genes
  - Non-coding genes (lncRNA, miRNA, etc.)
  - Regulatory sequences
- Extensive Data Describing Functionally Significant Variation
- Information about Population Genetic Structure of Research Colonies

Reference Genome Assembly for *Callithrix jacchus*

- 2010: Initial genome assembly and annotation
  - Published by Worley et al., Nature Genetics (2014)
  - DNA sample for reference genome from Southwest NPRC colony
  - Contig N50: 29.3 kb
- 2015: New assembly (Keio University)
  - Contig N50: 61.0 kb
- 2017: Another improvement (Broad Inst.)
  - Contig N50: 155.3 kb

Current “Best” assembly for common marmoset

- Assembly ASM275486v1 submitted by Broad Institute
- DNA sample from marmoset from New England NPRC

- Total sequence length: 2.845 gigabases
- Contig N50: 155.3 kb
- Scaffold N50: 129.2 Mb

RNA sequencing to define tissue-specific trranscriptome

- *Callithrix jacchus* – common marmoset

- 10 tissues (pituitary, spleen, lymph node, bone marrow, kidney, heart, skeletal muscle, liver, lung, colon)
- Illumina short read data produced by Baylor genome center
- Reads per tissue: 54.6 – 128.6 million
- Analysis: Chris Mason (Weill Cornell Med. Center)

Project Team: Nonhuman primate reference transcriptome project (Chris Mason, Michael Katze, Gary Schroth, Jeffrey Rogers)

Current “Best” assembly for common marmoset

- Assembly ASM275486v1 submitted by Broad Institute
- DNA sample from marmoset from New England NPRC

- Total sequence length: 2.845 gigabases
- Contig N50: 155.3 kb
- Scaffold N50: 129.2 Mb

Ensembl annotation

- Protein coding genes: 19,690
- Non-coding genes: 8,922

Coming soon (a few months?)

- New whole genome assembly for *Callithrix jacchus* in progress
- Evan Eichler (Univ. of Wash.) and Wes Warren (McDonnell Genome Institute, Wash. Univ.)
- De novo assembly using long read technologies, additional scaffolding
- Eichler-Warren gorilla assembly: 3.1 gigabases; Contig N50 9.6 Mb

Discovering Functionally Significant Genetic Variation in Marmosets

- Initial studies in 2010-2013: Whole genome sequences from seven animals

- Dr. Rosario: Whole genome sequences from >80 animals
- Several million single nucleotide variants identified

- The cost of whole genome sequencing continues to fall, approaching $1000 per genome

Why is discovering functional variation in marmosets important ?

- Discovery of novel “damaging” mutations can lead directly to new naturally occurring models of human genetic disease
- Knowledge of functional variation among marmosets can facilitate better selection of animals for specific experiments
- Information about functional variation among marmosets allows for more thorough interpretation of research results

Why is discovering functional variation in marmosets important ?

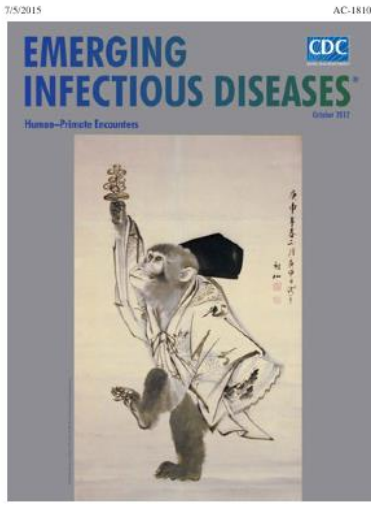
- Discovery of novel “damaging” mutations can lead directly to new naturally occurring models of human genetic disease
- Knowledge of functional variation among marmosets can facilitate better selection of animals for specific experiments
- Information about functional variation among marmosets allows for more thorough interpretation of research results

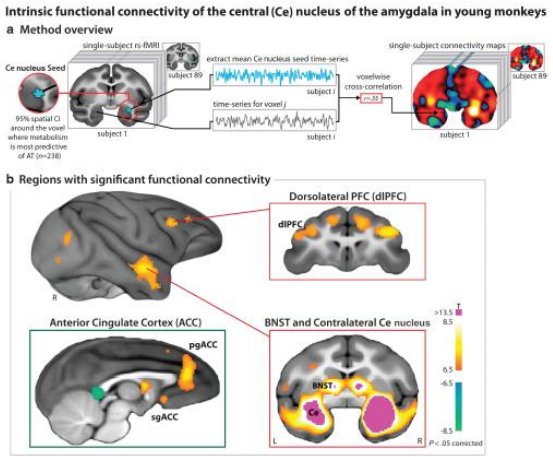
- Selection of animals for gene editing

How much genetic variation is segregating among rhesus macaques ?

How much of that variation is functionally significant ?

Can we use that variation to investigate questions related to either primate evolutionary adaptation or human health and disease ?





The population genomics of rhesus macaques (*Macaca mulatta*) based on whole-genome sequences

Cheng Xue,[1] Muthuswamy Raveendran,[1] R. Alan Harris,[1,2] Gloria L. Fawcett,[1,19] Xiaoming Liu,[3] Simon White,[1] Mahmoud Dahdouli,[1,20] David Rio Deiros,[1] Jennifer E. Below,[3] William Salerno,[1] Laura Cox,[4] Guoping Fan,[5] Betsy Ferguson,[6] Julie Horvath,[7,8,9] Zach Johnson,[10,21] Sree Kanthaswamy,[11,12] H. Michael Kubisch,[13] Dahai Liu,[14] Michael Platt,[15,16] David G. Smith,[11] Binghua Sun,[14] Eric J. Vallender,[13,17,22] Feng Wang,[2] Roger W. Wiseman,[18] Rui Chen,[1,2] Donna M. Muzny,[1] Richard A. Gibbs,[1,2] Fuli Yu,[1,2] and Jeffrey Rogers[1,2]

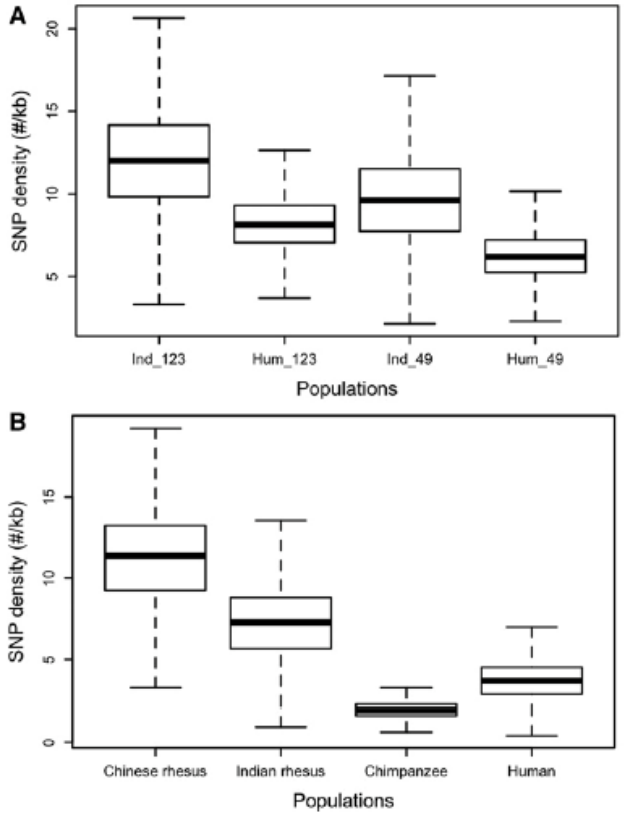
Genome Research 26:1651–1662
www.genome.org

HGSC
HUMAN GENOME SEQUENCING CENTER

Baylor College of Medicine

Discovery of Single Nucleotide Polymorphism in Rhesus Macaques

- n = 133 rhesus macaque whole genome sequences
- 124 Indian-origin = 31.9 million SNVs
  13.66 million private alleles
- 9 Chinese-origin = 30.1 million SNVs
  11.81 million private alleles
- Total Rhesus SNVs = 43.77 million



Analysis: R.A. Harris

Discovery of Single Nucleotide Polymorphism in Rhesus Macaques

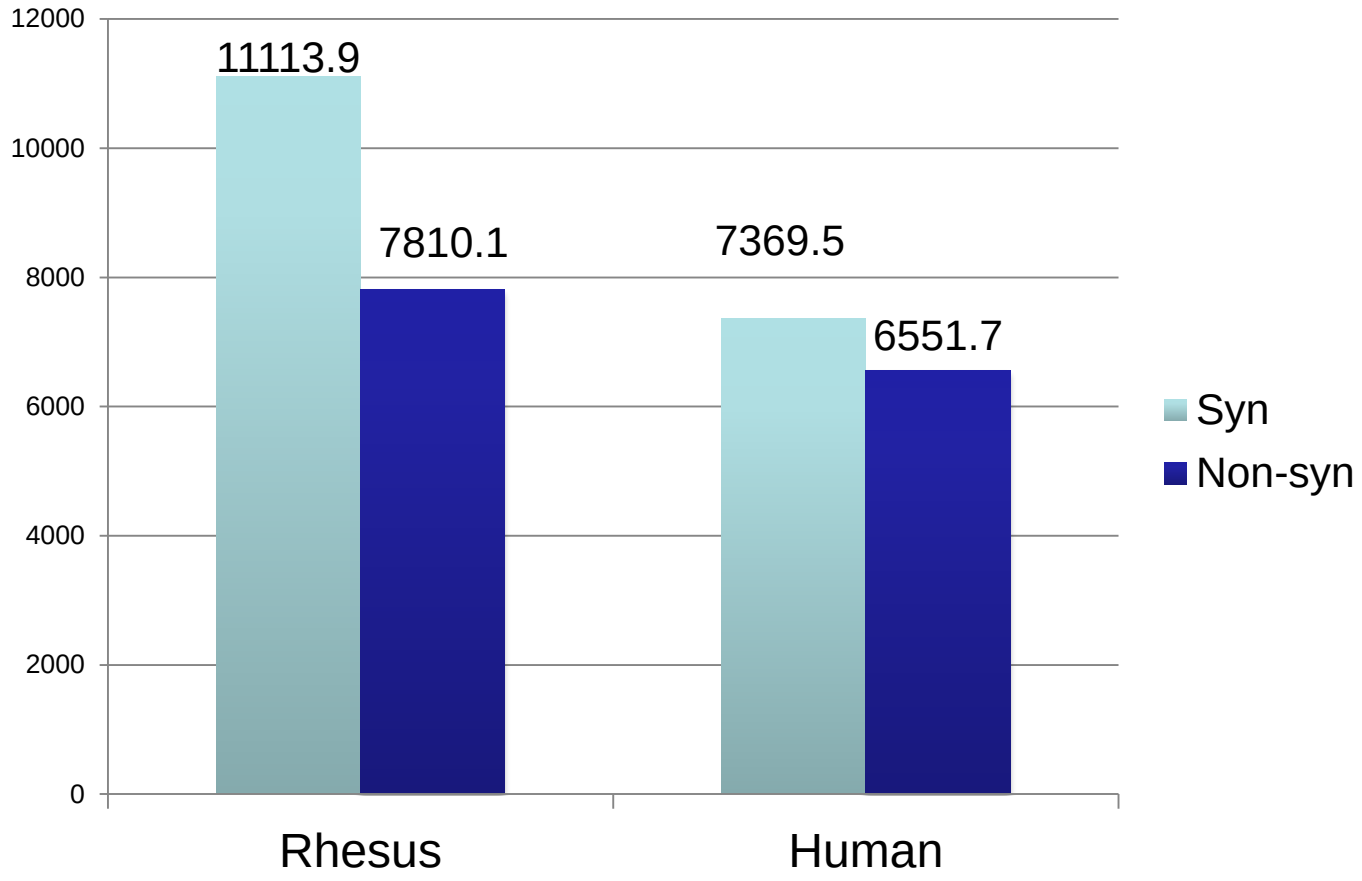
- n = 133 rhesus macaque whole genome sequences
- 124 Indian-origin = 31.9 million SNVs
  13.66 million private alleles
- 9 Chinese-origin = 30.1 million SNVs
  11.81 million private alleles
- Total Rhesus SNVs = 43.77 million

VEP prediction	Number of rhesus variants observed
Missense	126,445
Splice region	42,054
Stop codon gained	2,642
Mature miRNA	650

Analysis: R.A. Harris

Rhesus vs. Human: Nonsynonymous / Synonymous ratio

Average number of variants per individual



Sample size n=133

Whole Genome Sequencing data across 12 primate research colonies	Number of animals
Tulane National Primate Research Center	143
California National Primate Research Center	126
Wisconsin National Primate Research Center	96
Oregon National Primate Research Center	78
Caribbean Primate Research Center (CPRC), Cayo Santiago	35
The University of Texas MD Anderson Cancer Center, Bastrop	16
New England Primate Research Center	14
Southwest National Primate Research Center	6
Yerkes National Primate Research Center	7
Wild caught Chinese	3
Hazleton--Texas Primate Center	1
Labs of Virginia	1

Total sample size n = 526

SNV results from 526 rhesus macaques

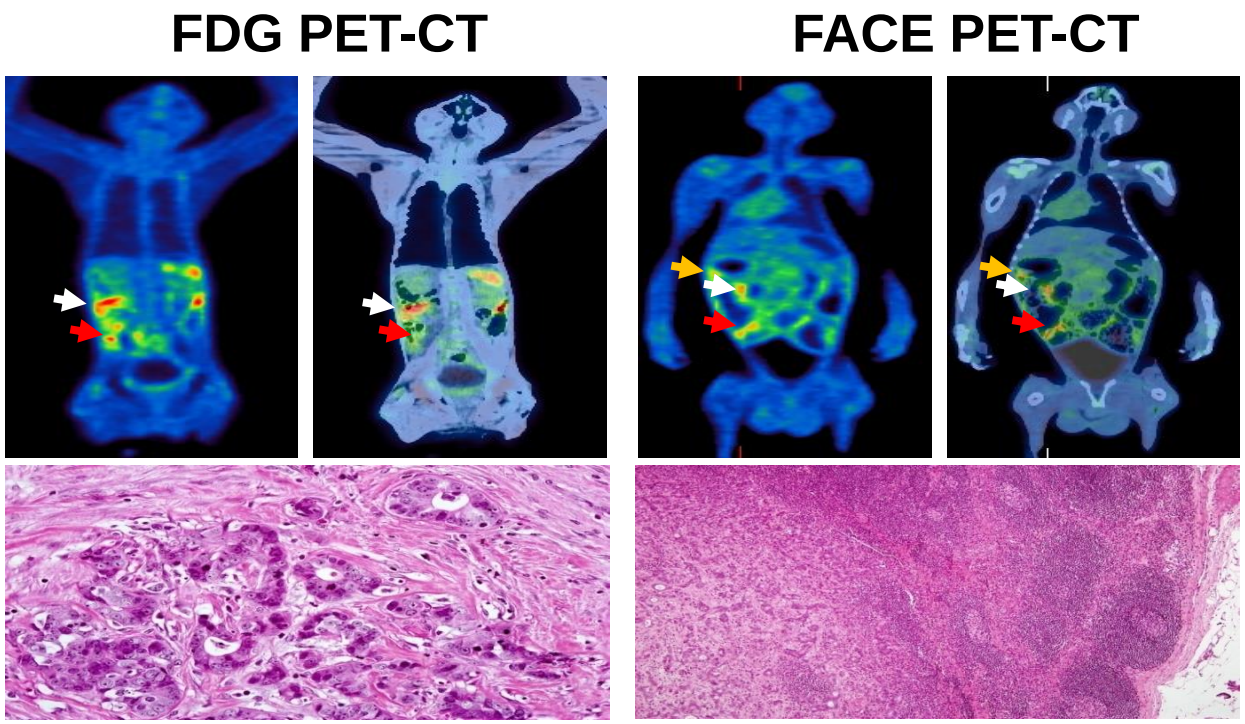
Total number of variant SNV sites identified	72,746,387
Number of singletons	17,616,218
Average number of SNVs per individual	9,476,124
Average heterozygosity	0.0020
Number of missense variants	340,104
Number of genes affected by missense variants	19,924
Number of de novo stop codons gained	8,556

We have observed missense mutations in 19,924 different genes:
94.4% of protein coding genes annotated in the rhesus genome

Lynch Syndrome

Hereditary Colon Cancer caused by mutations in DNA mis-match repair genes

Lynch Syndrome - autosomal dominant hereditary colorectal cancer



- Prevalence in humans of 1 in 440
- 2-7% of colorectal cancer cases

Beth Dray, DVM and Christian Abee, DVM
MD Anderson Keeling Center
for Comparative Medicine and Research
Bastrop, TX)

Lynch Syndrome

Hereditary Colon Cancer caused by mutations in DNA mis-match repair genes

Lynch Syndrome - autosomal dominant hereditary colorectal cancer

Gene	Frequency in affected patients
MSH2	60%
MLH1	30%
MSH6	7-10%
PSM2	Infrequent
PSM1	Case Report
TGFBR2	Case Report

- Prevalence of 1 in 440
- 2-7% of colorectal cancer cases

Lynch Syndrome: Candidate mutations in rhesus macaques

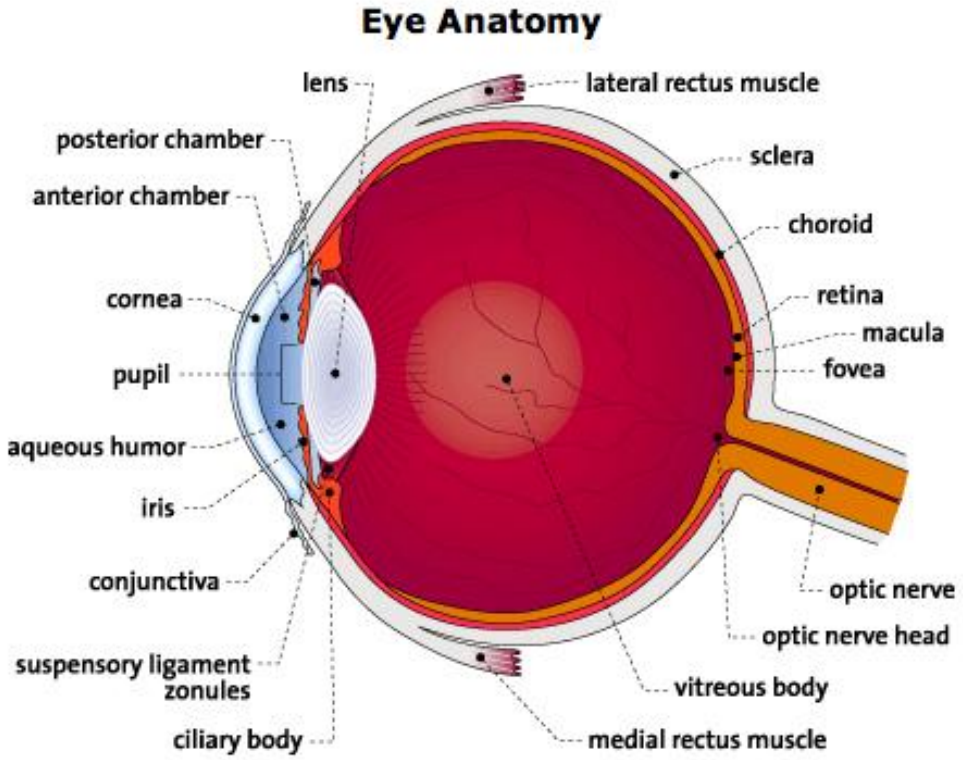
SNV	FaST-LMM p value	Variant Count (Het's)	Case MAF	Control MAF	VEP Consequence	Gene Symbol	CADD PHRED Score
chr13:48556451	1.9×10^{-25}	6 / 20	0.15	0	missense	*MSH6*	22.8
chr13:48564908	2.2×10^{-12}	6 / 20	0.15	0	downstream	*MSH6*	5.033
chr2:105789825	3.4×10^{-16}	4 / 20	0.10	0	stop gained	*MLH1*	36

- **N=20 rhesus macaques diagnosed with colorectal cancer**
- All rhesus SNVs lifted over to human coordinates
  - same reference base
  - same consequence except for chr13:48564908 in *MSH6* isoform 3' UTR
- CADD Score >= 20 means 1% most functionally significant SNPs in **human** genome

Dray et al. (2018)
Genes and Cancer
Vol 9: 142-152

Rhesus as models for human eye disease

Rui Chen, Jeffrey Rogers, Timothy Stout

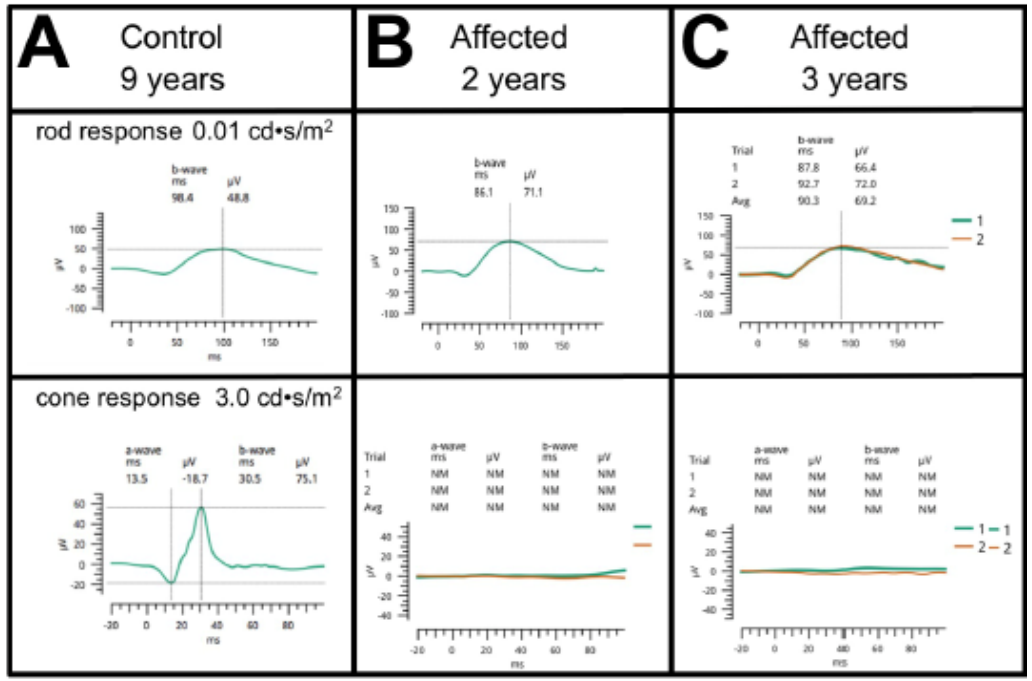


Do macaques carry damaging mutations in genes known to cause retinal diseases in humans ?

Discovery of new model of cone dystrophy

Sara Thomasy, Ala Moshiri, Jeff Roberts, Rui Chen, Tim Stout, Jeff Rogers

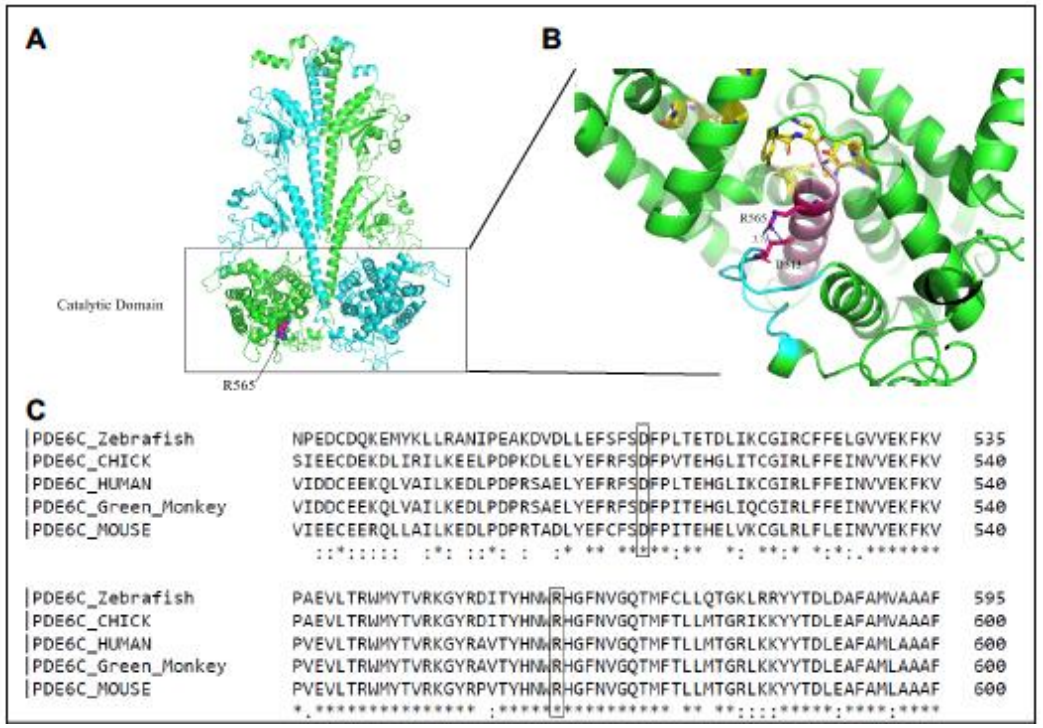
- Behavioral observations at California National Primate Res. Center suggested that two juvenile rhesus macaques had partial visual impairment
- Ophthalmic examination by Drs. Thomasy and Moshiri revealed near complete loss of cone photoreceptor function with normal rod photoreceptors



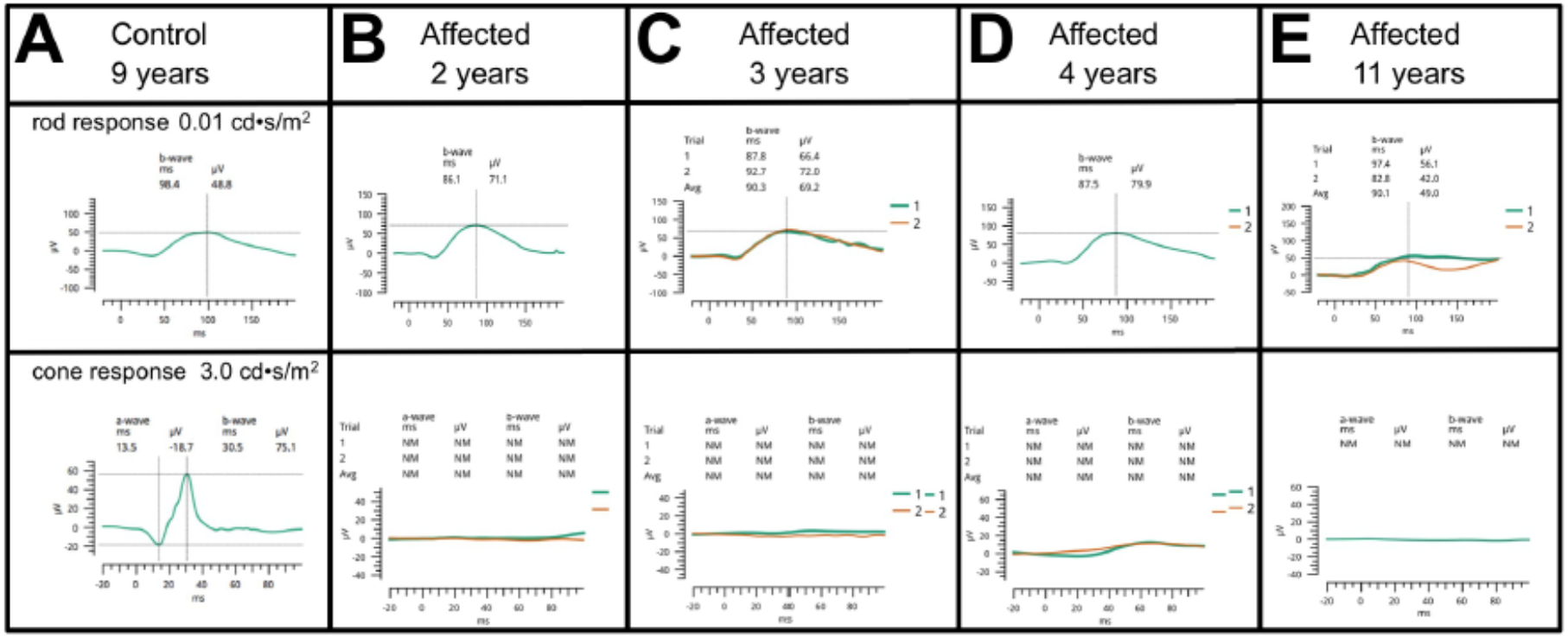
Moshiri et al. (in revision)

Discovery of new model of cone dystrophy

- Whole genome sequencing identified a missense mutation in *PDE6C* that is homozygous in both affected animals
- This gene codes for an enzyme that is expressed in cone photoreceptors and is critical to the phototransduction cascade. The enzyme hydrolyzes cGMP causing gated channels to close.
- *In vitro* functional assay shows this mutation essentially eliminates enzymatic function (N.O. Artemyev)



Moshiri et al. (in revision)



Moshiri et al. (in revision)

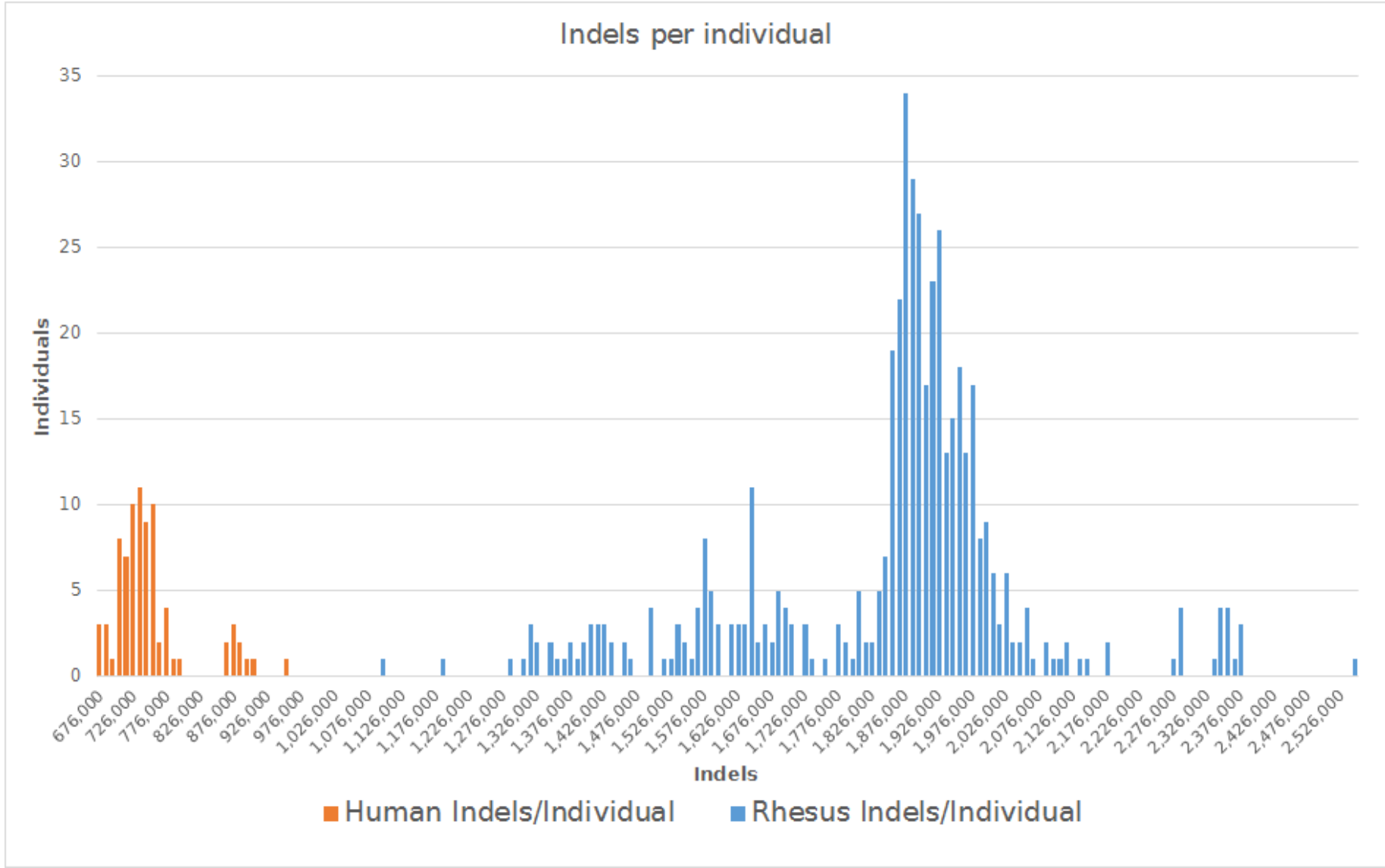
SNV results from 526 rhesus macaques

Total number of variant SNV sites identified	72,746,387
Number of singletons	17,616,218
Average number of SNVs per individual	9,476,124
Average heterozygosity	0.0020
Number of missense variants	340,104
Number of genes affected by missense variants	19,924
Number of de novo stop codons gained	8,556

We have observed missense mutations in 19,924 different genes:
94.4% of protein coding genes annotated in the rhesus genome

Expectations for functional variation in marmosets

- There is no reason to believe marmosets will have significantly lower levels of functionally significant genetic variation than rhesus macaques
- Functional variation can be the subject of investigation on its own and lead to new genetic models
- Functional variation can serve as modifying or compensatory variants when outbred animals are used for gene-editing experiments


Indels per individual
35
30
25
20
15
10
5
0
Individuals
676,000
726,000
776,000
826,000
876,000
926,000
976,000
1,026,000
1,076,000
1,126,000
1,176,000
1,226,000
1,276,000
1,326,000
1,376,000
1,426,000
1,476,000
1,526,000
1,576,000
1,626,000
1,676,000
1,726,000
1,776,000
1,826,000
1,876,000
1,926,000
1,976,000
2,026,000
2,076,000
2,126,000
2,176,000
2,226,000
2,276,000
2,326,000
2,376,000
2,426,000
2,476,000
2,526,000
Indels
Human Indels/Individual
Rhesus Indels/Individual

Fundamental Resources for Primate Genomics

- Whole Genome Reference Assembly
  - Essentially complete (few gaps) with minimal sequence errors
- Accurate and “Complete” Annotation
  - Protein coding genes
  - Non-coding genes (lncRNA, miRNA, etc.)
  - Regulatory sequences
- Extensive Data Describing Functionally Significant Variation
- Information about Population Genetic Structure of Research Colonies

Acknowledgments

Baylor College of Medicine

Human Genome Sequencing Ctr.

- Muthuswamy Raveendran
- R. Alan Harris
- Donna Muzny
- Richard Gibbs
- Many staff contributors

Keeling Center for Comparative Medicine

MD Anderson Cancer Center

- Beth Dray
- Christian Abee

California National Primate Research Center

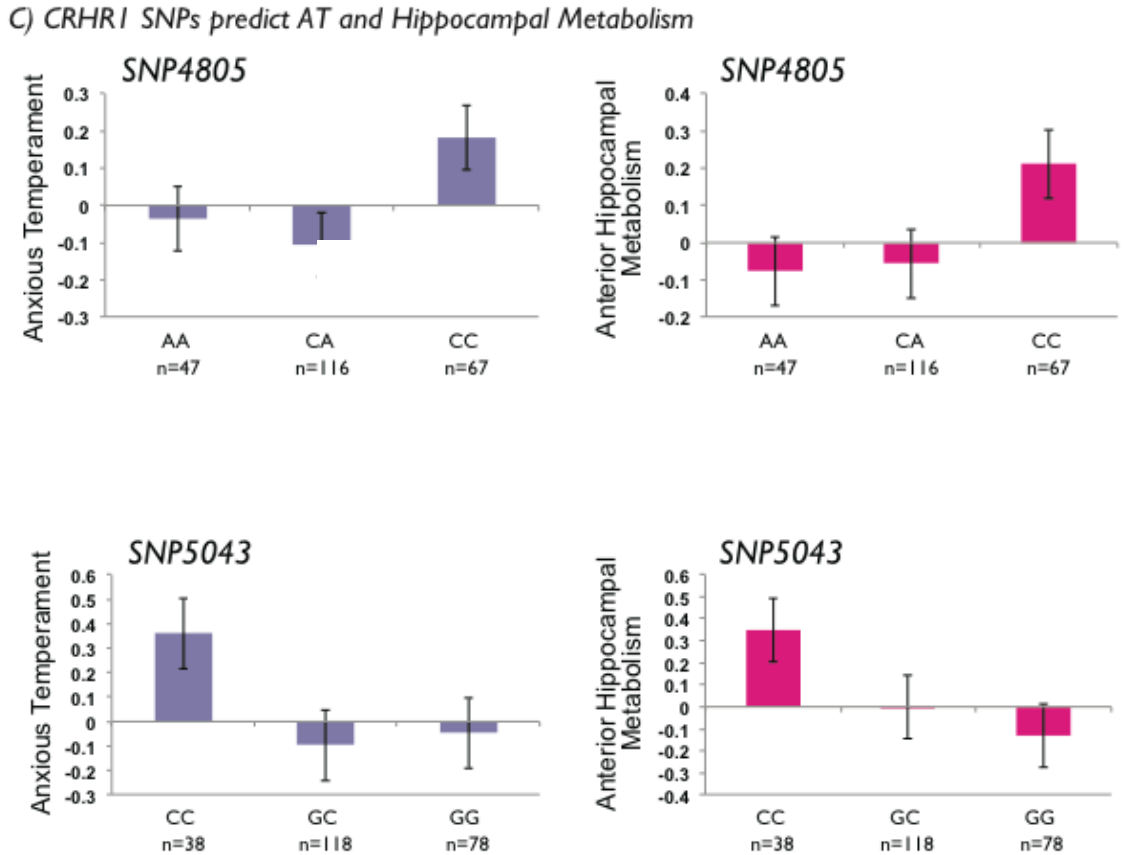
- Sara Thomasy
- Ala Moshiri
- Jeff Roberts

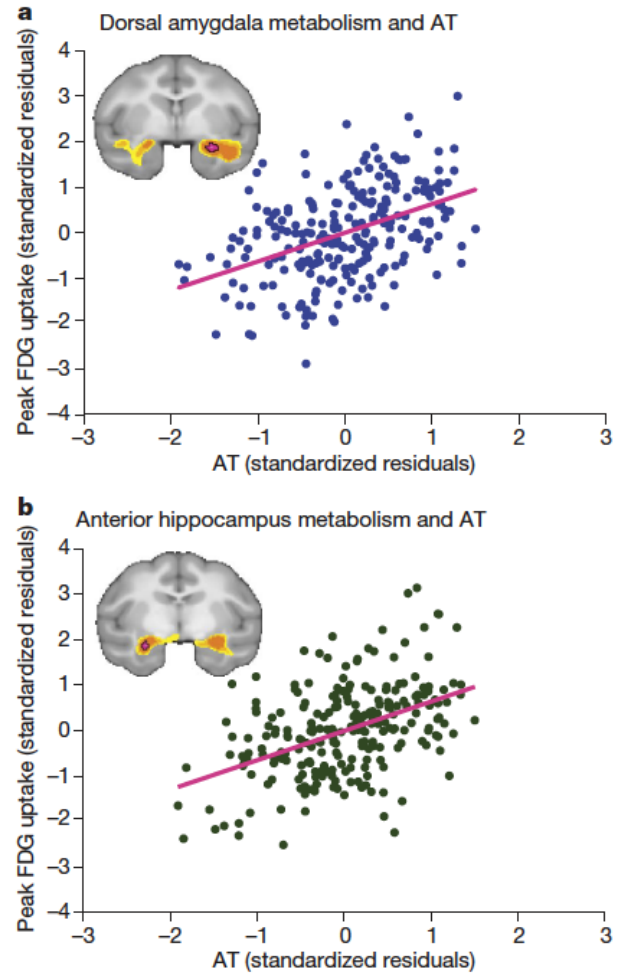
Southwest National Primate Research Center

- Suzette Tardif

Funding

- NIH Office of Research Infrastructure Programs
- NIH National Human Genome Institute





CRHR1 genotypes, neural circuits and the diathesis for anxiety and depression

J Rogers[1,2], M Raveendran[1], GL Fawcett[1], AS Fox[3,4], SE Shelton[5,6], JA Oler[5,6], J Cheverud[7], DM Muzny[1], RA Gibbs[1], RJ Davidson[3,4,5,6] and NH Kalin[3,4,5,6]

Molecular Psychiatry (2013) 18, 700–707