Mutation and sequence analysis

Jeffrey Chuang
The Jackson Laboratory for Genomic Medicine

May 2018

The Jackson Laboratory
Leading the search for tomorrow's cures

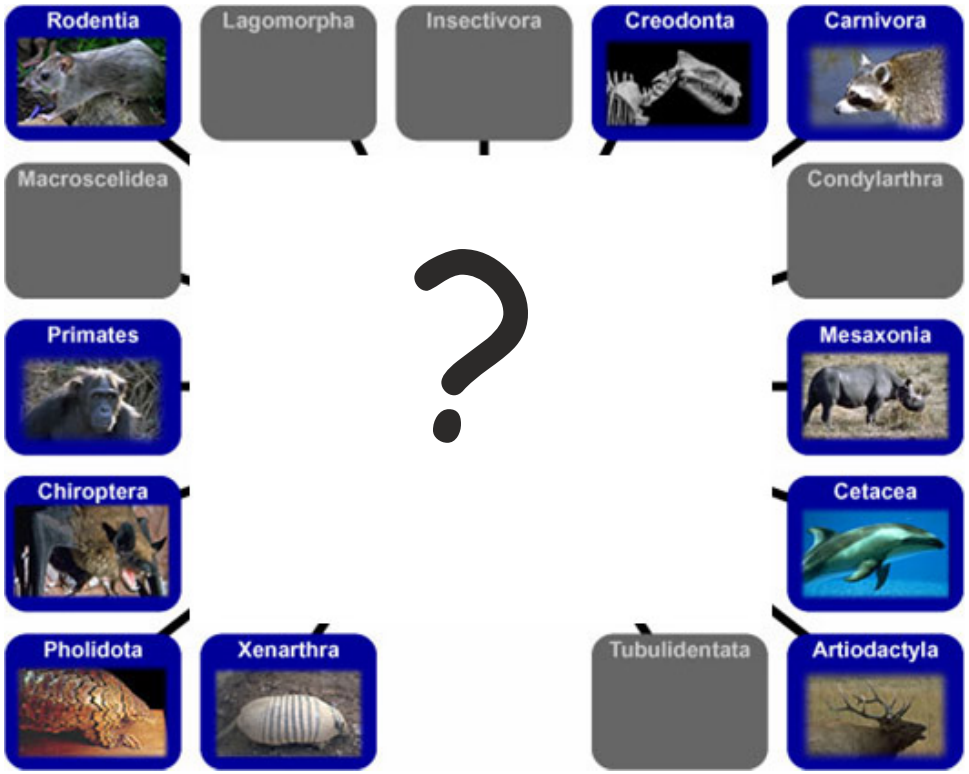
Mutations and Evolution

”Nothing in biology makes sense, except in the light of evolution.

Without that light it becomes a pile of sundry facts - some of them interesting or curious but making no meaningful picture as a whole”

-Theodosius Dobzhansky

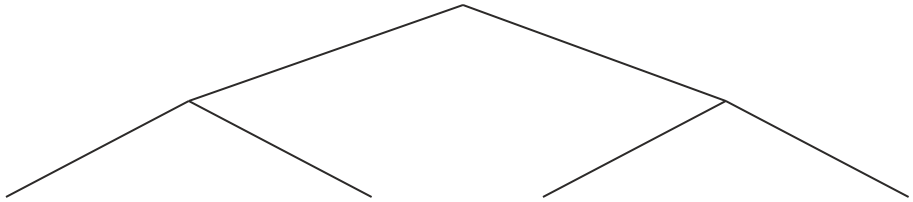
Classification as a motivation for sequence analysis



How should different species be understood? If they are related, then in what way?

Pre-genomic classification

- **Hierarchical classification**
(Carolus Linnaeus, 1707-1778)



- **Binomial nomenclature**

Name Pre-Linnaeus: Apis pubescens thorace subgriseo abdomine fusco pedibus posticis glabris utrinque margine ciliatis.

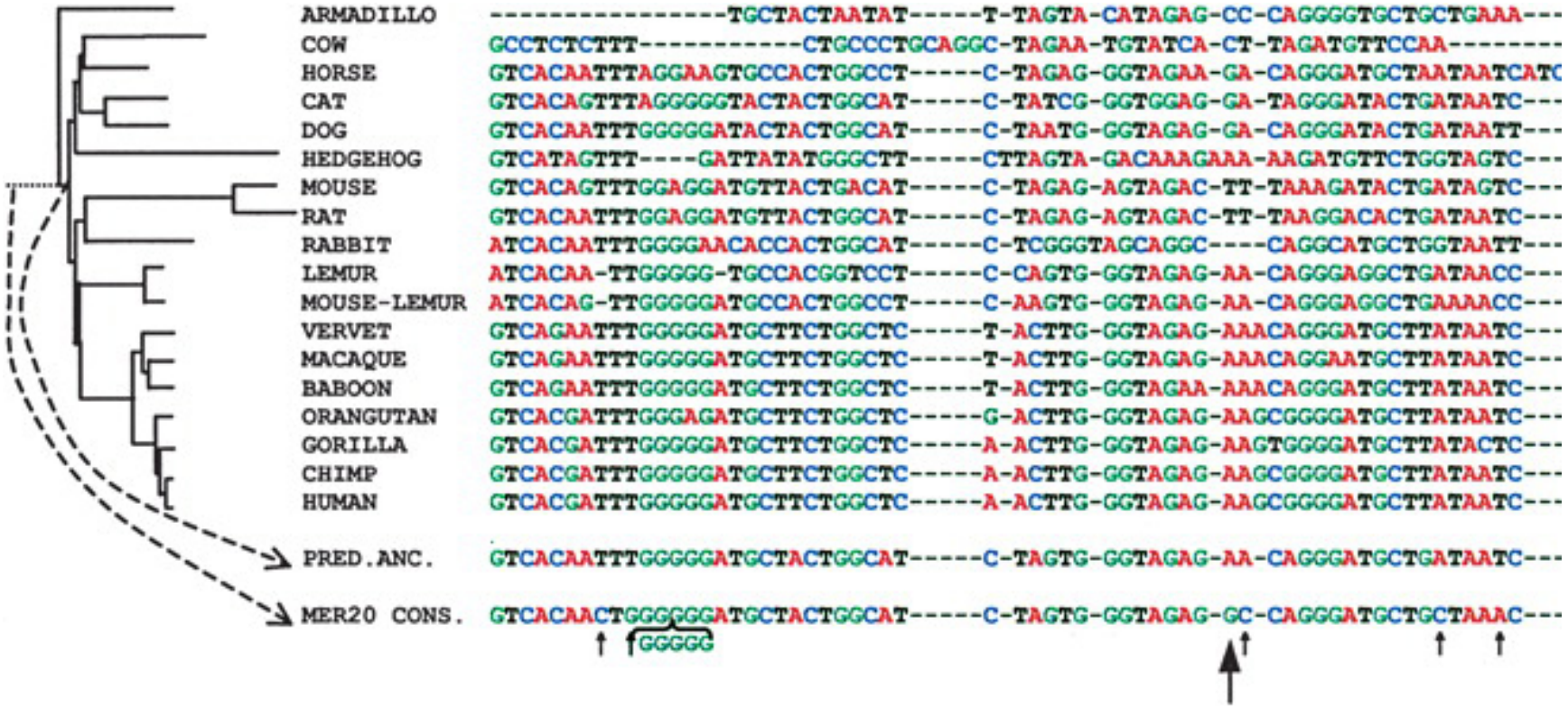
Name Post-linnaeus: Apis mellifera.

Genomic characters for classification

Traditional classification based on *arbitrary* morphological characters.
Example: number and arrangement of plant reproductive organs.

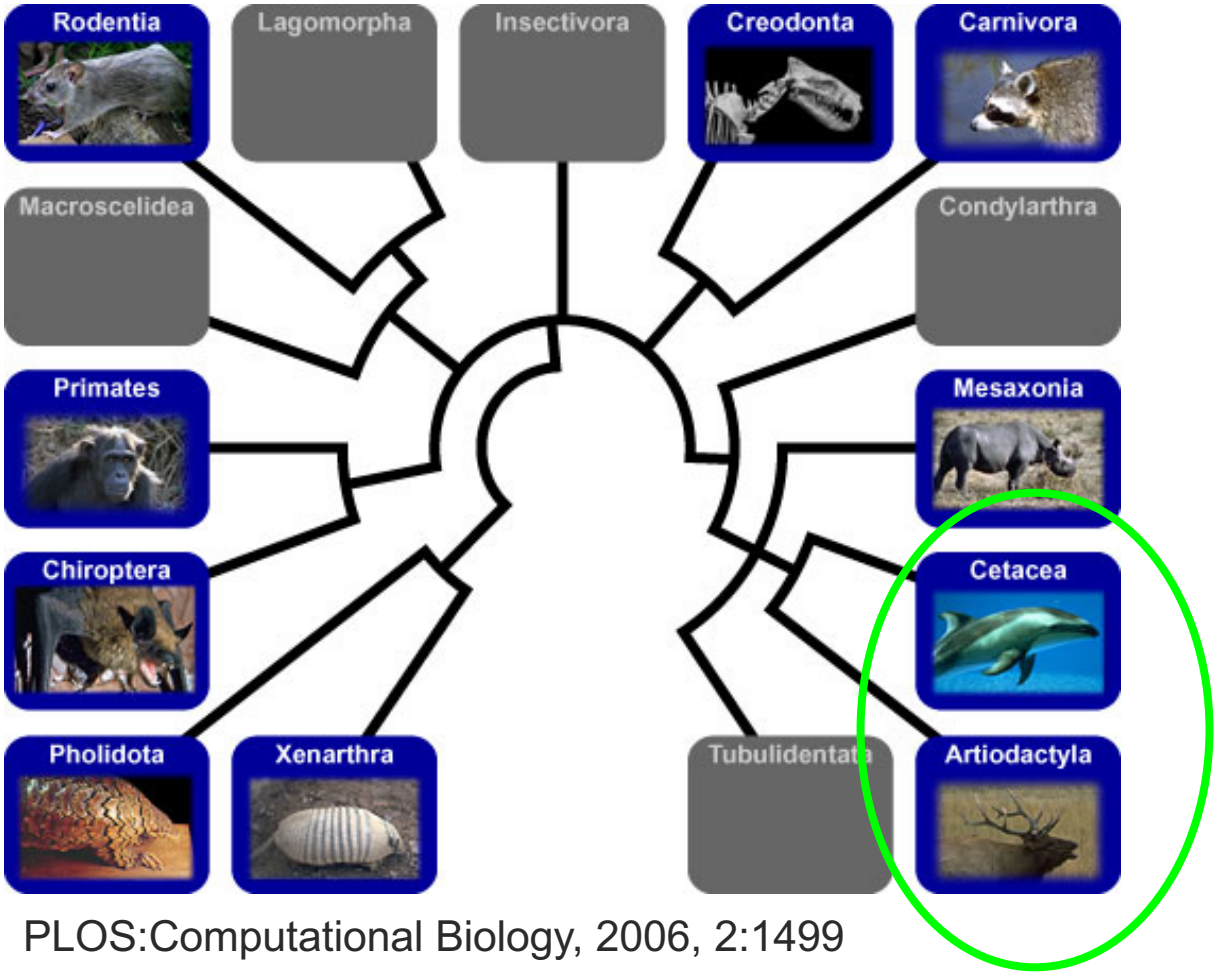
Molecular sequence data is universal (A, C, G, T), digital (4 nts), and abundant (millions to billions of bases per genome)

Modern classification is based on sequence evolution



Mathieu Blanchette et al. Genome Res. 2004; 14: 2412-2423

Modern classification is based on sequence evolution



PLOS:Computational Biology, 2006, 2:1499

Sequence Comparison

Cross-species analysis

1) Assemble genomes,
2) Align regions of interest
3) Count differences across species

Intraspecies analysis

1) Align reads to reference,
2) Count differences to reference

To compare samples, repeat steps (1) and (2) for each sample.

Sequence comparison leads naturally to topics on:

Phylogenetics

Evolutionary analysis from high-throughput sequencing data

Computational methods

Phylogenetics

Traditional, well-covered topic.

- Tree-building methods
- Distance metrics
- Species evolution

The Jackson Laboratory

Leading the search for tomorrow's cures

Evolutionary sequence analysis

The Jackson Laboratory

Leading the search for tomorrow's cures

The 1000 Genomes Project

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

Table 1 | Summary of 1000 Genomes Project phase I data

	Autosomes	Chromosome X	GENCODE regions*
Samples	1,092	1,092	1,092
Total raw bases (Gb)	19,049	804	327
Mean mapped depth (×)	5.1	3.9	80.3
SNPs			
No. sites overall	36.7 M	1.3 M	498 K

1000G project identified 1 SNP per 100 nt. 5x whole genome sequencing per individual; 80x in gene regions.

1000 Genomes Project Consortium *et al. Nature* **491,** 56–65 (2012).

Human population evolution

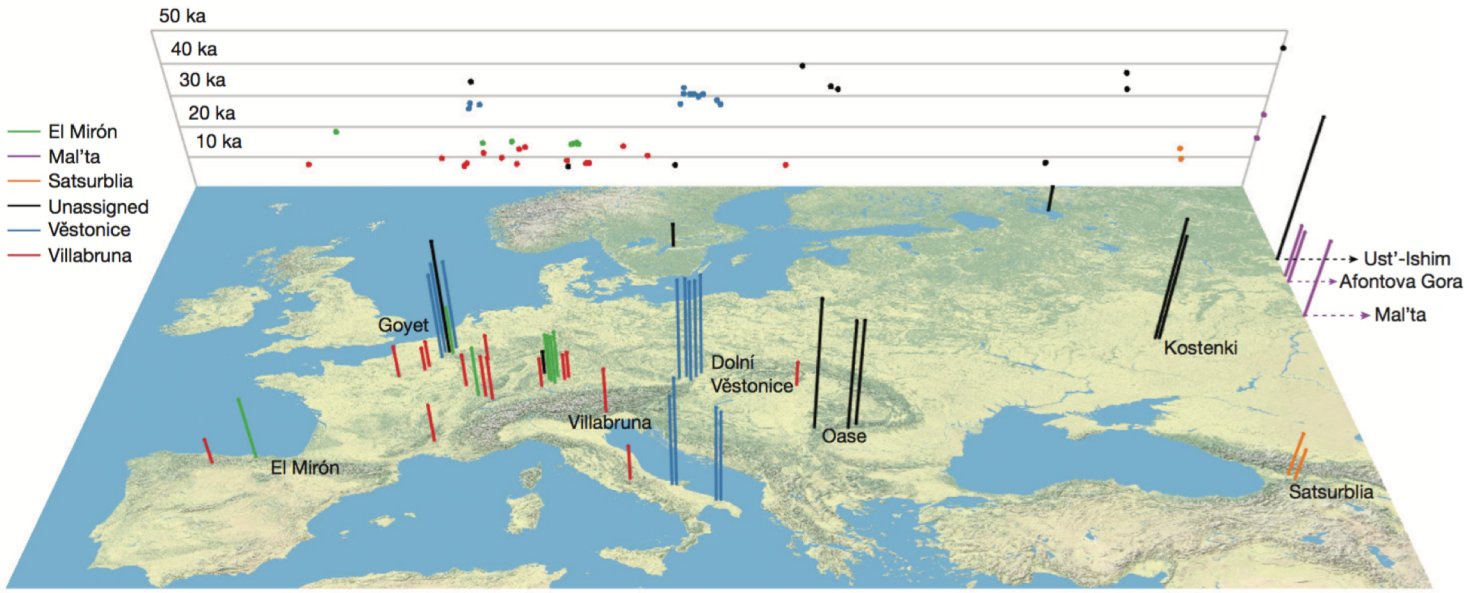


Figure 1 | Location and age of the 51 ancient modern humans. Each bar corresponds to an individual, the colour code designates the genetically defined cluster of individuals, and the height is proportional to age (the background grid shows a projection of longitude against age). To help in visualization, we add jitter for sites with multiple individuals from nearby locations. Four individuals from Siberia are plotted at the far eastern edge of the map. ka, thousand years ago.

Analysis of polymorphism data allows for detailed reconstruction of population history

Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* (2016). doi:10.1038/nature17993

The Cancer Genome Atlas

ARTICLE

doi:10.1038/nature11404

Comprehensive genomic characterization of squamous cell lung cancers

The Cancer Genome Atlas Research Network*

ARTICLE

OPEN
doi:10.1038/nature12222

Comprehensive molecular characterization of clear cell renal cell carcinoma

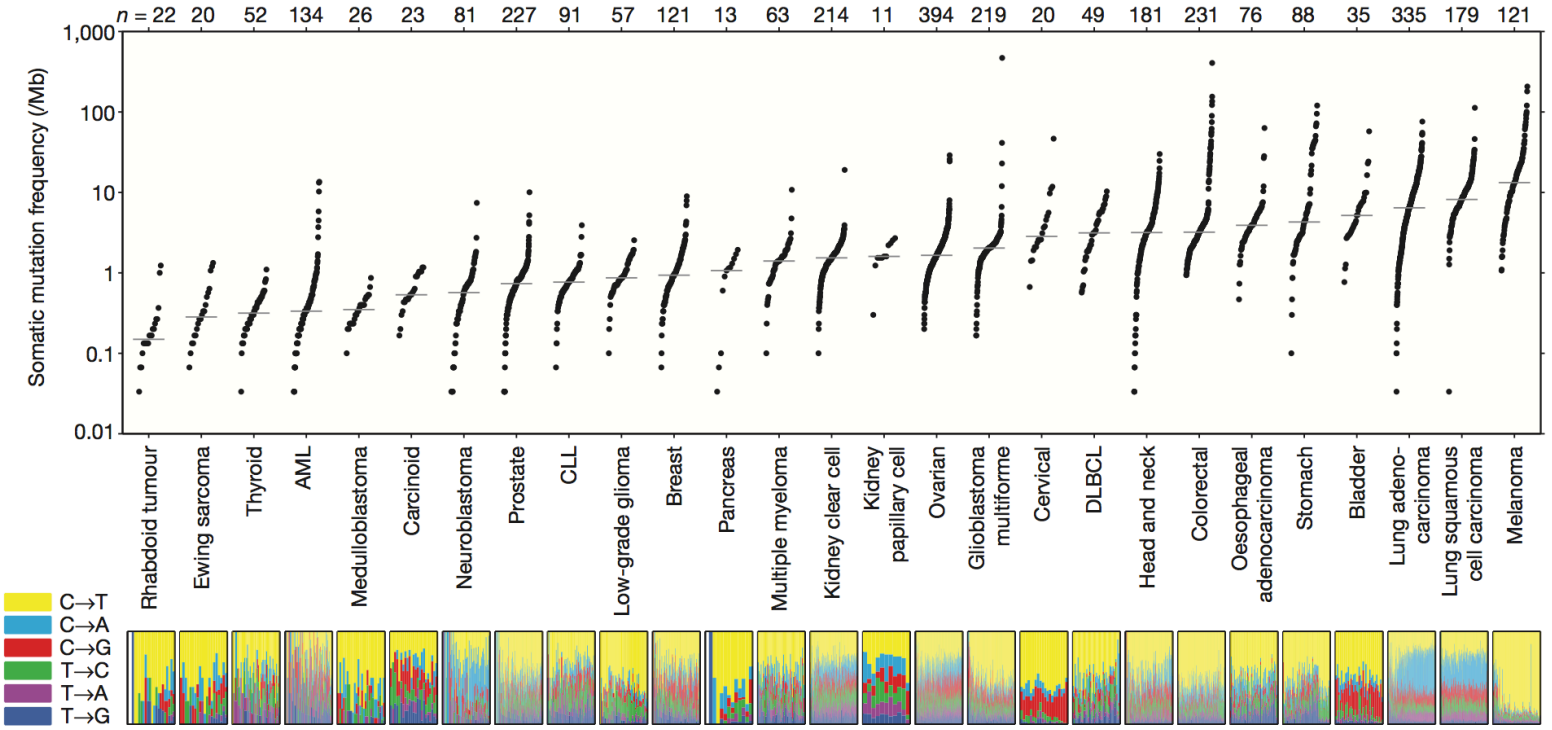
The Cancer Genome Atlas Research Network*

ARTICLE

Comprehensive molecular portraits of human breast tumours

The Cancer Genome Atlas Network*

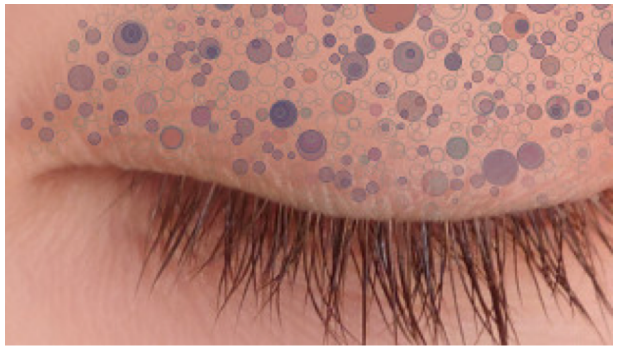
Mutations in cancer



Mutation rates vary across different cancer types

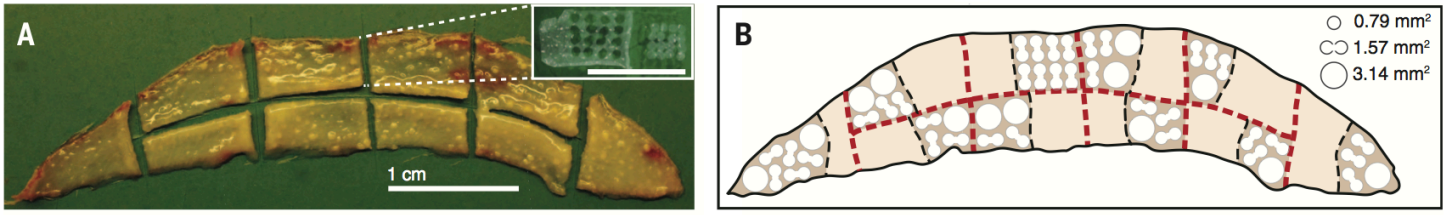
Lawrence, M. S. *et al. Nature* **499,** 214–218 (2013).

Example: high-depth sequencing of multiple regions of eyelid skin



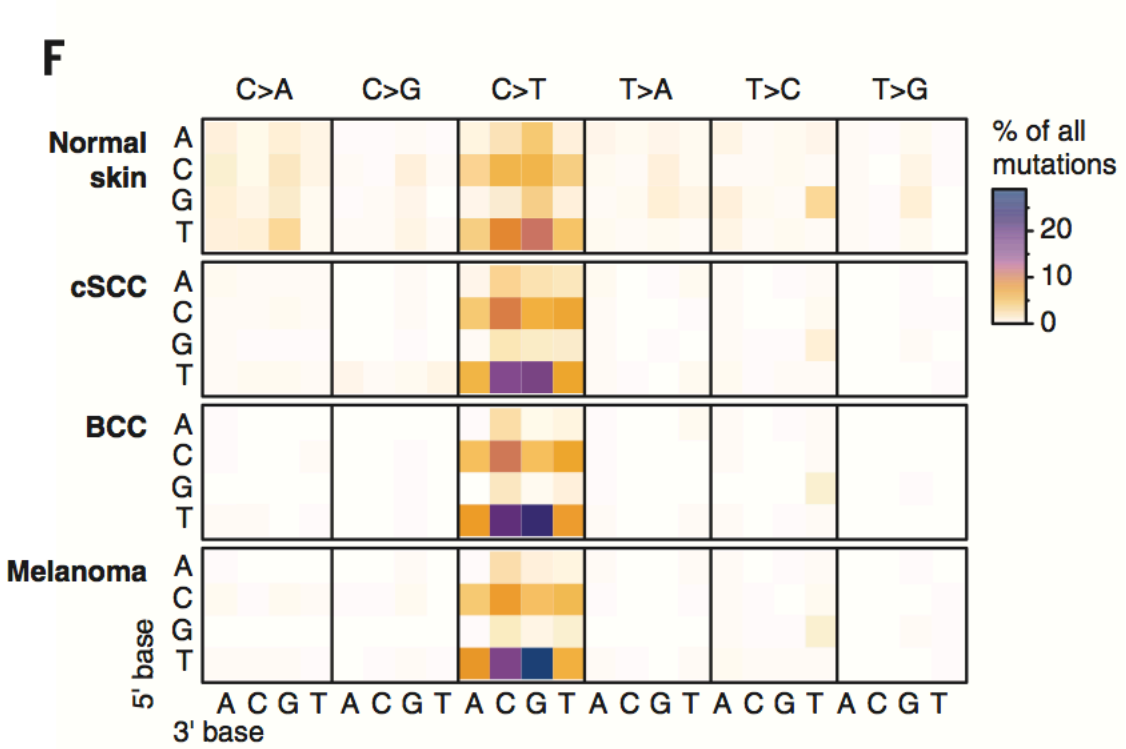
Representation of mutant clones in normal eyelid skin.

Martincorena et al. Science 348:6237 (2015).



Excised human eyelid obtained via cosmetic surgery.

Mutations are similar across normal skin and skin cancers



Heat map of rates of each mutation type, depending on the nucleotides upstream and downstream of the mutated base.
SCC = squamous cell carcinoma. BCC = basal cell carcinoma

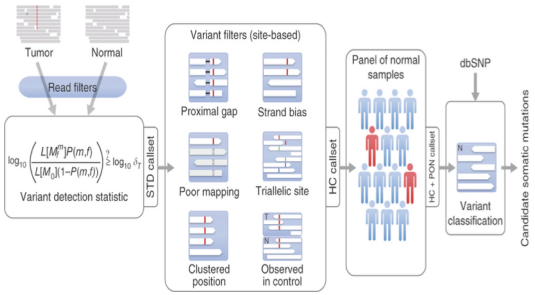
Computational Methods

The Jackson Laboratory

Leading the search for tomorrow's cures

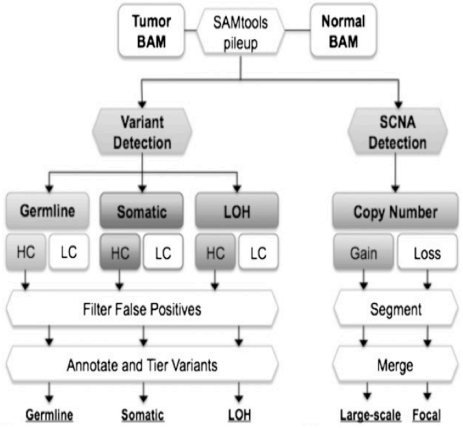
Examples of Mutation Callers

MuTect



Cibulskis et al. Nature Biotechnology (2013)

Varscan



Koboldt et al. Genome Research (2012).

SomaticSniper

$$S = -10\log_{10}\left(\sum_{G_i=0}^{9}\frac{P(T|G_i)P(G_i)P(N|G_i)P(G_i)}{\sum_{G_j=0}^{9}P(T|G_j)P(G_j)\sum_{G_k=0}^{9}P(N|G_k)P(G_k)}\right)$$

Larson et al. Bioinformatics (2011).

There are a variety of computational pipelines that compute mutations from sequencing data.
Broader resources: Genome Analysis Toolkit (GATK), GenomeSpace, Galaxy, KnowEng, et al.

Variant Call Format files

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 27 no. 15 2011, pages 2156–2158*
doi:10.1093/bioinformatics/btr330

Sequence analysis Advance Access publication June 7, 2011

The variant call format and VCFtools

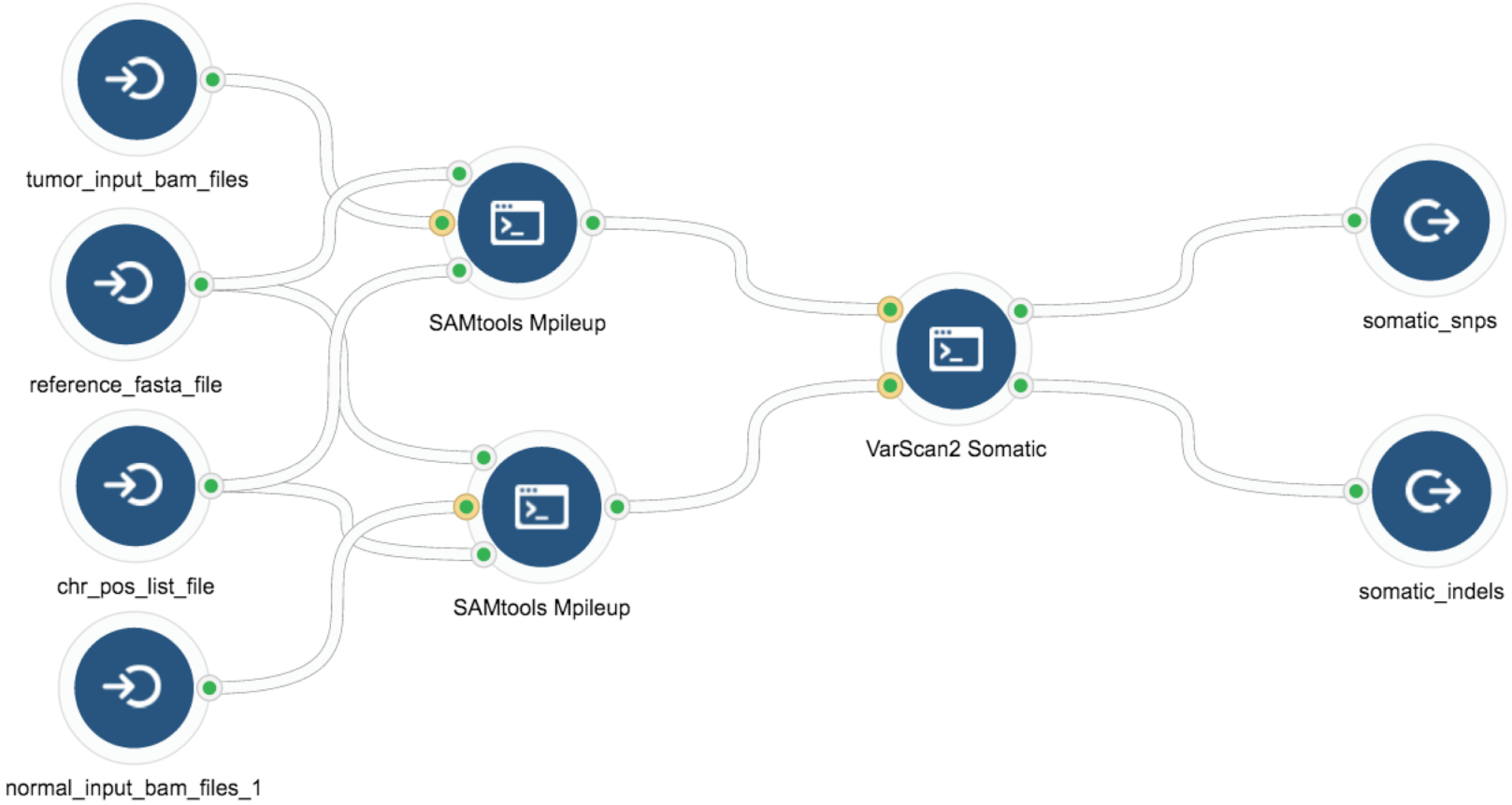
Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Associate Editor: John Quackenbush

Polymorphism data is stored in VCF files.

Example mutation calling pipeline in graphical workflow



Seven Bridges Genomics

THE JACKSON LABORATORY

Exercises

The Jackson Laboratory

Leading the search for tomorrow's cures

Driver mutation exercise

Download file “cancer driver data.xls” on Canvas website.

These are mutation data from 3 tumors (ME050, ME100L, ME024) of the same type, and they share a common driver mutation. Try to identify the common driver.

Don’t look up this reference until you have attempted to identify the driver. Nature 485, 502–506 (24 May 2012) doi:10.1038/nature11071

Driver mutation exercise solution

Copy the column **Protein_Change** to a text file named test.txt.

Open R

Read the file into R: data=read.csv(“test.txt”)

Find mutations that appear repeatedly: sort(table(data))

TCGA cancer exercises

Go to http://www.cbioportal.org/

Verify the mutation from the prior example by analyzing the Melanoma Broad/Dana Farber, Nature 2012 dataset. View the mutations in this set by going to the Data Sets tab in cBioPortal and selecting this set.

What are the 4 most commonly mutated genes in this dataset?

Why do you think the top 3 genes are more commonly mutated than the one you found in the Driver Mutation exercise? Useful link: Ensembl.http://www.ensembl.org/Homo_sapiens/Info/Index

What is the most commonly mutated gene in breast cancer? What is the typical diagnosis age? Use data from the Breast Invasive Carcinoma project. 1002 cases (TCGA Cell 2015).

Example pipeline exercise

Identify and validate coding variants from exome sequencing data

http://recipes.genomespace.org/view/17

This is a good example for mutation calling using the program FreeBayes and data publicly available on GenomeSpace.

We won't execute this in class because data transfer and compute times are too long. This would be a good homework assignment for students.