Analysing *Alu* inserts detected from high-throughput sequencing data

Harun Mustafa

Mentor: Matei David
Supervisor: Michael Brudno

July 3, 2013

Computational Biology Lab
Computer Science
UNIVERSITY OF TORONTO

NSERC CRSNG

Before we begin...

Even though I'll only present the minimal amount of biology required to understand my presentation, neglecting some fundamental concepts, I still have to introduce a lot of terms and concepts. So please **feel free to interrupt and ask questions**. It will try my best to avoid leaving you guys feeling like this:

Overview

1. Biology background
   - 1.1 Human genetic diversity
   - 1.2 Molecular biology
   - 1.3 Genomes, the human genome
   - 1.4 High-throughput sequencing and genome assembly
   - 1.5 *Alu* sequences, novel *Alu*s
2. Novel *Alu* insert sequence assembly
   - 2.1 Assembly pipeline
   - 2.2 Applications
     - 2.2.1 Tracing origins of the inserts
     - 2.2.2 Subfamily clustering (not discussed today)
3. Use of novel *Alu*s for measuring genetic distance

Biological background: Human genetic diversity

- *Homo sapiens* originated in Africa
  - Oldest remains found in Ethiopia
- Small group migrated to Arabia, then spread
- Confirmed by much higher genetic diversity between African populations
  - i.e. On average, different populations are much less related to each other
- Roughly inverse relation between distance to Africa and interpopulation genetic diversity

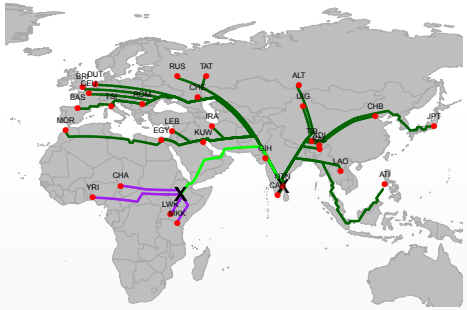
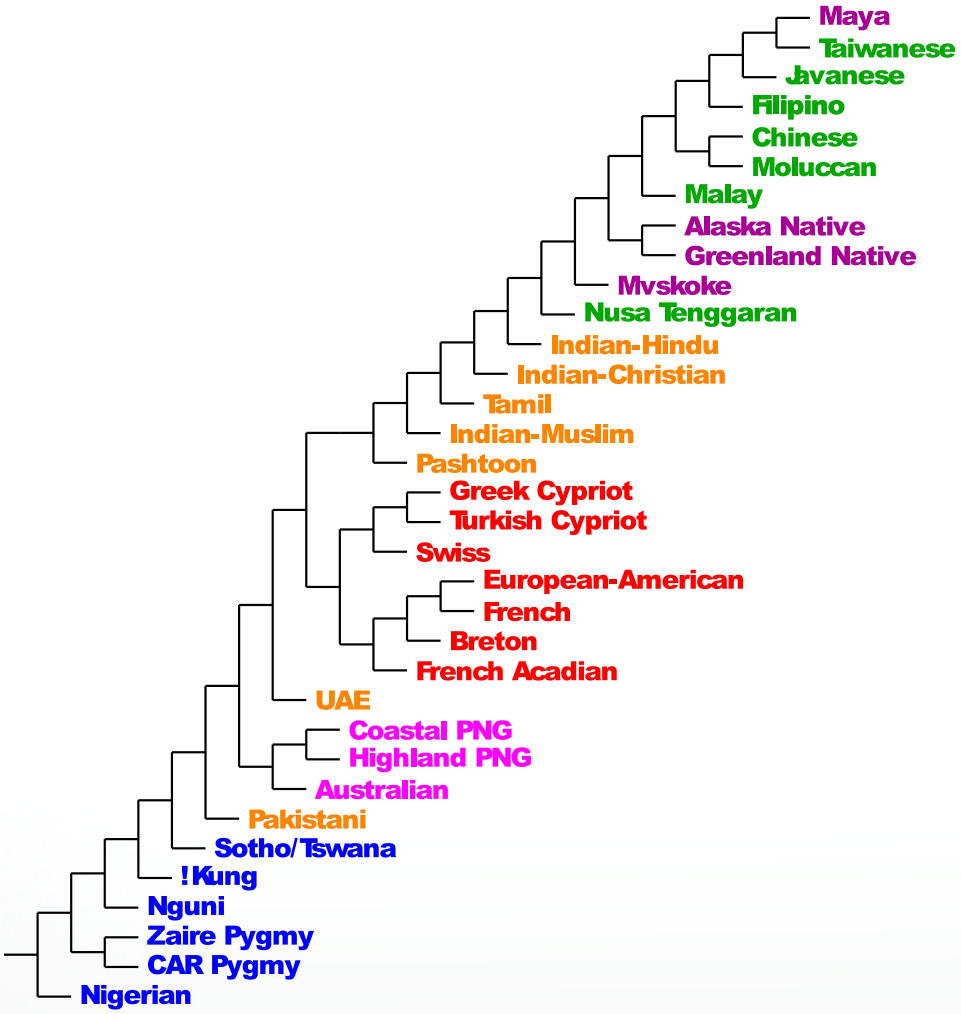


Figure: Melé, Marta, et al. "Recombination gives a new insight in the effective population size and the history of the Old World human populations." Molecular biology and evolution 29.1 (2012): 25-30.

Biological background: Human genetic diversity



Stoneking, Mark, et al. "Aluinsertion polymorphisms and human evolution: Evidence for a larger population size in Africa." Genome research 7.11 (1997): 1061-1071.

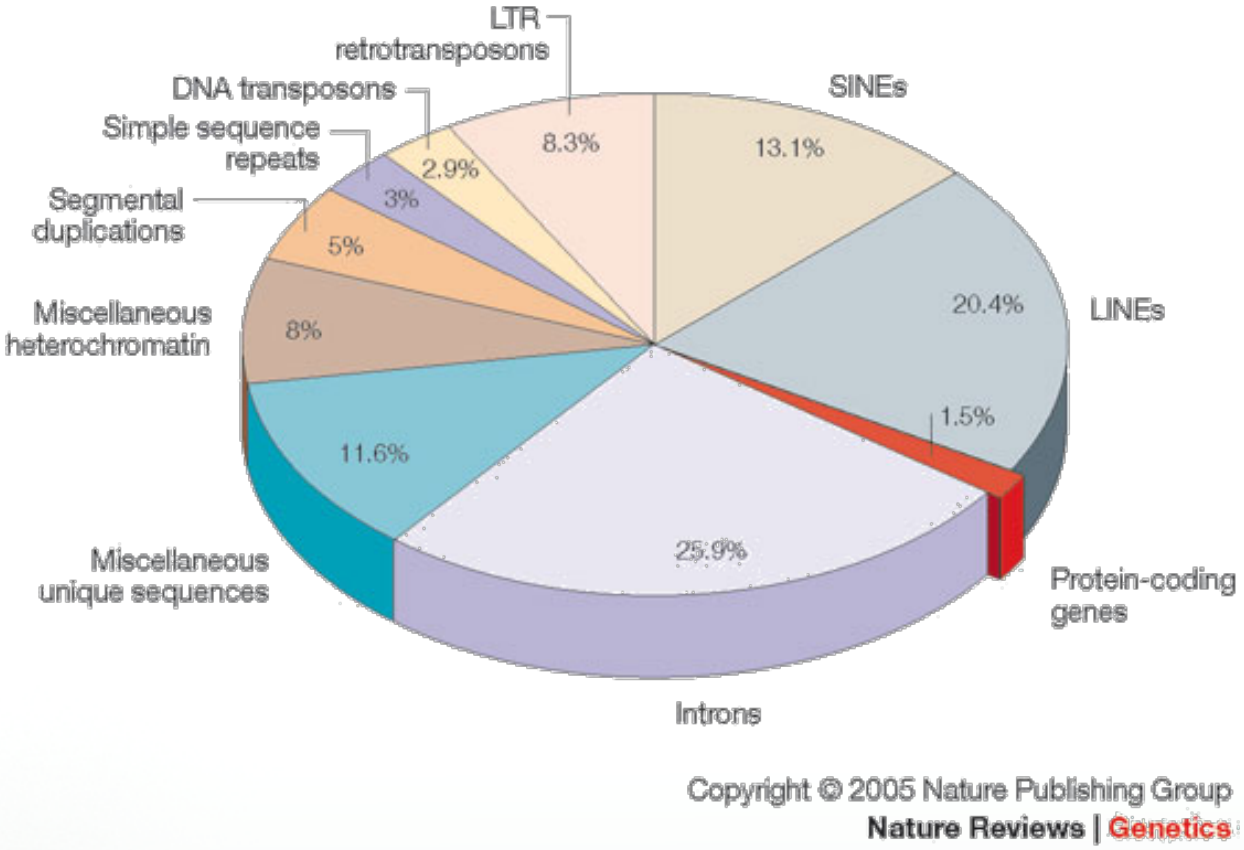
Biological background: Molecular Biology

- Central dogma DNA $\longrightarrow$ RNA $\longrightarrow$ Protein
  - Only in general, several significant exceptions exist
  - DNA stores almost-read-only information. Think of it as a sequence of genes (words) and gene regulators connected by long lengths of junk that may or may not affect these regulators
  - RNA made by transcribing small section of the DNA, can be processed
  - Proteins (and sometimes RNA) perform biological functions
- Genome is collection of chromosomes (DNA molecules) in a cell
  - eg. Humans have two copies of 23 chromosomes (one from each parent)
- When cell divides, DNA replicated so each daughter cell gets a copy
  - Basis of genome sequencing (reading)

Biological background: Genomes

- A chromosome can be represented as a sequence (string) in a five base pair (bp) alphabet $\Sigma = \{A, T, G, C, N\}$
  - Biologically only the first 4 exist, N represents an unknown
  - Represent substrings with coordinates
    - eg. `chr1:12345-12533`
  - This encoding does not represent structure, and other important information
- Consensus sequences for each chromosome from a couple of individuals serves as a reference genome
  - In human reference, variation among contributors not represented
  - An individual can be concisely represented as a list of variants (differences) between him/her and the reference

Biological background: Human genome

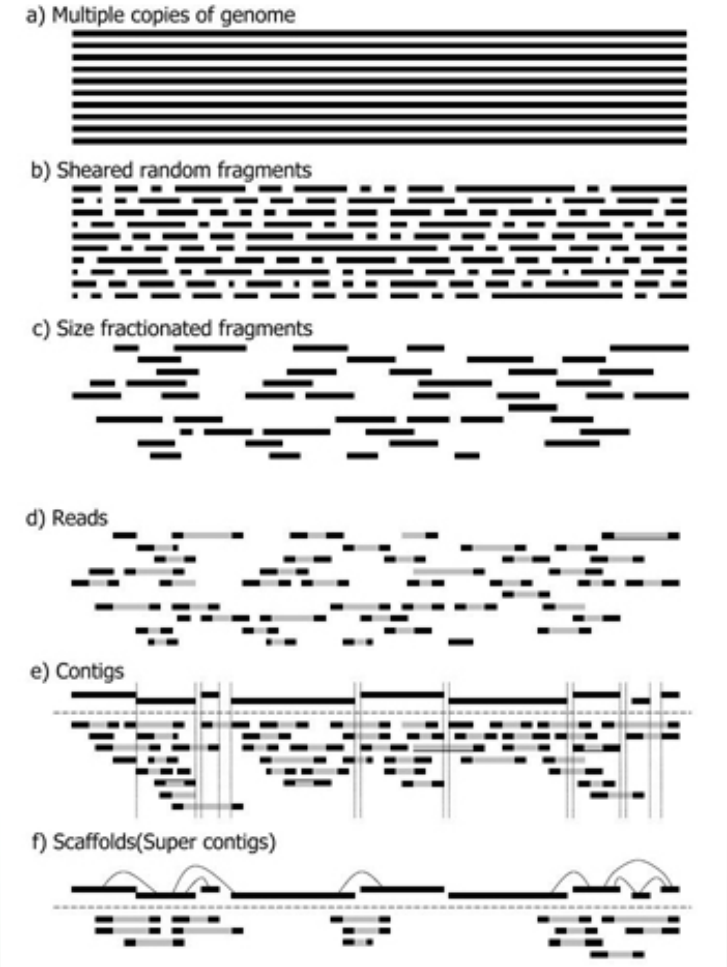


Gregory, T. Ryan. "Synergy between sequence and size in large-scale genomics." Nature Reviews Genetics 6.9 (2005): 699-708.

Biological background: High-throughput sequencing; genome assembly

- A chromosome can't be sequenced in one reaction
- Instead, molecules replicated several times, broken into tiny <100bp pieces, sequenced in parallel to get reads
  - Average number of reads per position called coverage
  - Coverage variable
- Reassemble the original chromosome strings by overlapping the reads
  - If reference sequence available, can align reads to it to help
- Several issues make this non-trivial
  - Don't know which chromosome a given read comes from
  - Repeats make it hard to figure out which reads overlap and how
  - etc.
- Several techniques and algorithms developed

Biological background: High-throughput sequencing; genome assembly



http://www.k.u-tokyo.ac.jp/pros-e/person/shinichi_morishita/shinichi_morishita.htm

QUESTIONS?

Biological background: *Alus*

ggccgggcgcggtggctcacgcctgtaatcccagcactttgggaggccgaggcgggcgg

- ∼45% of human genome composed of repetitive elements
  - Retrotransposons (LINEs, SINEs), other short repeats, etc.
- *Alus* are most common SINE, >1 million copies
- Replicate by reverse transcribing RNA to DNA then inserting into genome, 1 replication every ∼10 births
- ∼99% inactive in humans, *Alu*Y subfamily active
- Since lack of *Alu* rarely means that insertion didn't happen, good for calculating genetic distance
- Have been linked to certain disorders
- Traditionally detected through wet lab methods
- *Alu* inserts that are not present in the reference human genome called novel inserts

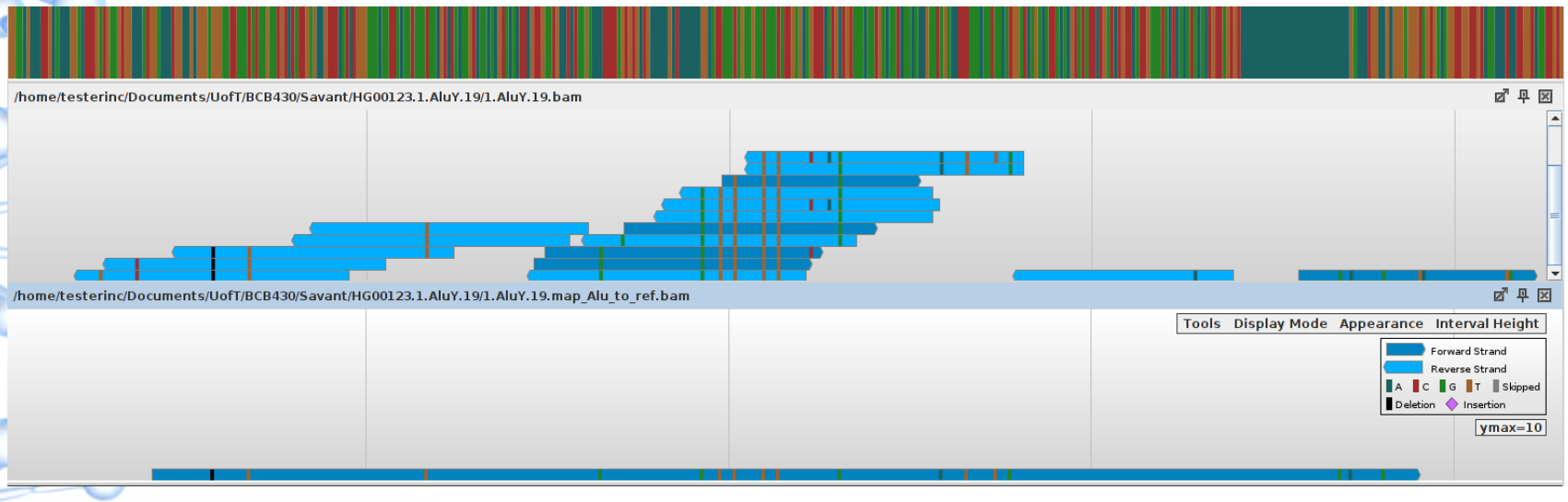
Biological background: Detecting novel *Alu*s

- Recently, programs developed for detecting novel *Alu* inserts from HTS data
  - Last year, Matei and I developed `alu-detect`
- All of these programs suffer from high false negative rates
  - Happens to be that coverage around *Alu*s is naturally low
  - Also dependent on result filtering
- Previous programs (including `alu-detect`) have only outputted lists of coordinates and subfamily identities of detected insertion events, not the actual sequences of the inserts

Assembling *Alu*s: Pipeline

- To assemble the sequence of a novel *Alu* insert, we need the reads
- `alu-detect` extended to report which reads used as evidence for novel inserts
- Pipeline of standard computational biology tools used to assemble the sequence
  - Aligned the reads to the consensus sequence of its *Alu* subfamily
  - Used read quality and alignment quality scores to determine significant differences between the reference and the reads
  - Merged the changes into the reference
- Since coverage was too low (not enough reads to make differences significant), data from several individuals was combined

Assembling *Alus*: Pipeline



Assembling *Alu*s: Origins of the inserts

- Aligned the constructed inserts back to the reference
- Since *Alu* sequences are very similar and have high copy number, several hits for each query
- Selected hit with highest alignment score
- Results need to be investigated further
  - High proportion of inserts originate from single source
  - Most common *Alu* subfamily in data is the previously reported most active
  - However, second most common subfamily in data not the second most active subfamily
  - Possible cause(s)

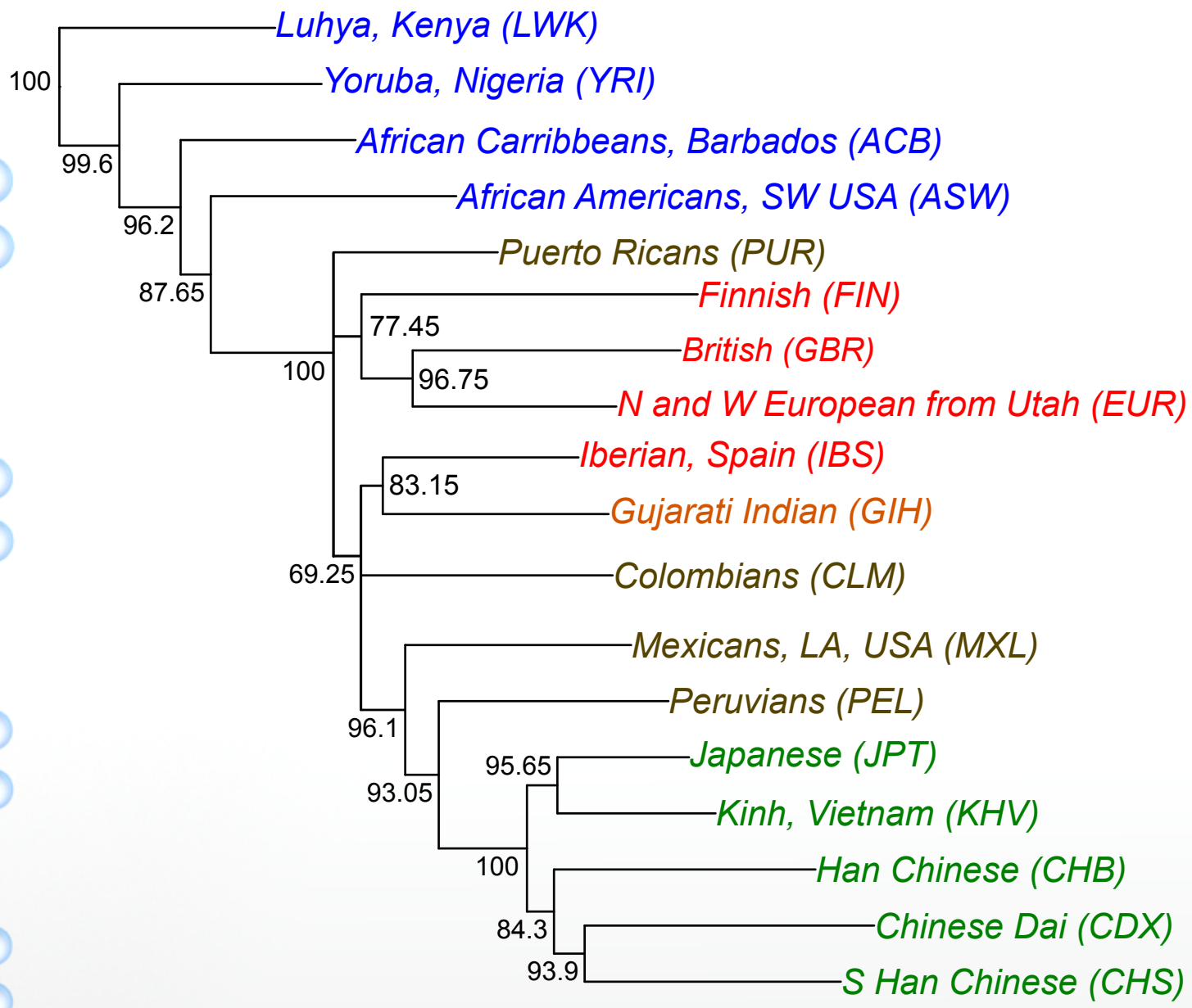
$$\left.\begin{array}{c}\text{Flawed method?}\\ \text{Garbage in, garbage out?}\end{array}\right\}\text{most likely}$$

Actual result?

Alu frequencies as genetic distance

- Detected novel *Alu* inserts in 166 individuals distributed across 19 ethnic groups from the 1000 Genomes project
- For each insert, calculated frequency of that insert in each group
  - Each insert represented as a 19-dimensional vector of probabilities
  - Equivalently, each population represented by a 50-dimensional vector of values in $[0, 1]$.
- Picked the top 50 inserts with the highest frequency variances
  - PCA also attempted, but results were much worse
- Clustered using neighbour-joining with the Cavalli-Szforza distance and bootstrapped with $B = 1000$ samples.
- Future work
  - See if these vectors are correlated to vectors of frequencies for other kinds of genetic variants

Population clustering by genetic distance



Thank you for your time

and now...

COOKIES!