Studying differences in mutability between parental sets of chromosomes

Hypothesis

Different genomes (i.e. sets of chromosomes) differ at many positions in sequence and genomic structure. Both characteristics, especially the latter (chromatin state), are known to impact mutation rates in our tissues, but how much of a difference does it make between parental sets of chromosomes?

Aim

We want to quantify global and local mutation rates of parental allele sets separately.

Methodology

First, we will review the literature on the matter and search for trios of parents+child who have been genome sequenced and for whom the child has had a tumour removed and sequenced (either genome or exome). This will also provide a direct way to answer our main question.

In the main part of the project, we will also infer parental origin without having the parents sequenced using an original approach:

1. We will identify patients of mixed-ethnicity in the non-TCGA ICGC cohort of cancer whole genomes (https://dcc.icgc.org/pcawg). The SNPs identified from germline genomes of these patients have been pre-phased using Beagle 5.1 (https://faculty.washington.edu/browning/beagle/old.beagle.html).
2. In parallel, we will simulate mixed-ethnicity genomes using the 1000G project (https://www.internationalgenome.org/) phased SNP data and benchmark probabilistic assignments of haplotypes to ethnicity.
3. We will apply our assignment of haplotypes to ethnicity to the identified mixed-ethnicity genomes.
4. We will phase pre-called somatic mutations (indels and SNVs) to SNPs (i.e. by using reads carrying both a SNP and the mutation) and thereby annotate them for the ethnicity.
5. We will quantify the local and global mutation rates per ethnicity/parent and compare them.

Theoretical basis of the project

Genome: The human genome is a complete set of nucleic acid sequences for humans, encoded as DNA within the 23 chromosome pairs in cell nuclei.

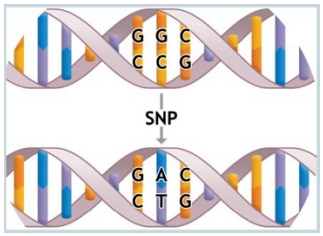
Chromosomes: Chromosomes are thread-like structures located inside the nucleus of animal and plant cells

ADN : deoxyribonucleic acid

Gene : A gene is the basic physical and functional unit of heredity. Genes are made up of DNA. Some genes act as instructions to make molecules called proteins. However, many genes do not code for proteins.

Allele : Refers to genes at the same level on the chromosomes of the same pair. An allele is one of two or more versions of DNA sequence (a single base or a segment of bases) at a given genomic location. An individual inherits two alleles, one from each parent, for any given genomic location where such variation exists.

SNPs :

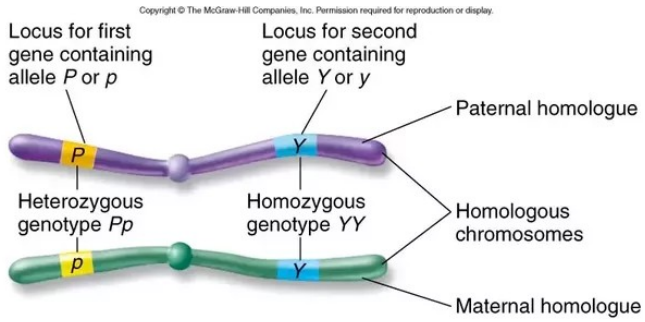


SNP : single nucleotide polymorphiosm -> is a single nucleotide associated polymorphism, a single base mutation in DNA. SNPs are "conserved" in the genome and represent the simplest form and the most common source of genetic polymorphism in the human genome.

Two randomly selected humans, regardless of geographic or ethnic origin, are 99.9% genetically identical so 0.1% explains genetic differences and biological variability.

Only 5% of the 2.85 billion A, C, G and T nucleotides that make up our genome are "protein expressed", which means that most genetic differences between two individuals are of no biological consequence.

Allele frequency :



Allele frequency represents the incidence of a gene variant in a population. Alleles are variants of a gene located at the same position, or genetic locus, on a chromosome. An allele frequency is calculated by dividing the number of times the allele of interest is observed in a population by the total number of copies of all alleles at that particular genetic locus in the population. Allele frequencies can be represented as a decimal, percentage or fraction. In a population, allele frequencies are a reflection of genetic diversity. Changes in allele frequency over time may indicate that genetic drift is occurring or that new mutations have been introduced into the population

Explanation 1000 genomes

The 1000 Genomes Project is a consortium of research scientists from around the world who have come together to use DNA sequencing to examine human genetic variation and find where our genomes differ. If you take two people, our DNA sequences DNA are very long. There are three billion bases, and they are almost identical to each other.

But there is a difference of about one place in a thousand. This means that there are approximately three million differences between each pair of people. The people who were selected to constitute a sample within the framework of the 1000 genomes project are anonymous people and selected so that they are fully representative of their population.

Initially, a polite phase was set up and in this phase a lot of genetic variations could be found: 15 million variants were found

The information from the project confirmed some of the things we knew about human evolution.

- For example, that we were born in Africa quite recently and developed from there.
- He taught us that natural selection has influenced virtually every part of our genome.

On top of that, we now have a catalog of a few thousand genes.

which we believe have been specifically and positively selected in our fairly recent history.

What we are working towards is a catalog, is a sort of encyclopedia that you can search for things.

So it's kind of a fundamental or basic framework for genetic research.

It took a lot of challenges to make the project work and get as far as we made it.

There's been a lot of innovation both in the technology – the machines to get the DNA sequence. but also in data processing and how we analyze and interpret genetic data.

Phase 3 and phase 1

The phase 1 variants list released in 2012 and the phase 3 variants list released in 2014 overlap but phase 3 is not a complete superset of phase 1. The variant positions between phase 3 and phase 1 releases were compared using their positions. This shows that 2.3M phase 1 sites are not present in phase 3. Of the 2.3M sites, 1.92M are SNPs, the rest are either indels or structural variations (SVs).

The difference between the two lists can be explained by a number of different reasons.

1. Some phase 1 samples were not used in phase 3 for various reasons. If a sample was not part of phase 3, variants private to this sample are not be part of the phase 3 set.
2. Our input sequence data is different. In phase 1 we had a mixture of both read lengths 36bp to >100bp and a mixture of sequencing platforms, Illumina, ABI

SOLiD and 454. In phase 3 we only used data from the Illumina sequencing platform and we only used read lengths of 70bp+. We believe that these calls are higher quality, and that variants excluded this way were probably not real.

3. The first two reasons listed explain 548k missing SNPs, leaving 1.37M SNPs still to be explained.

   The phase 1 and phase 3 variant calling pipelines are different. Phase 3 had an expanded set of variant callers, used haplotype aware variant callers and variant callers that used de novo assembly. It considered low coverage and exome sequence together rather than independently. Our genotype calling was also different using ShapeIt2 and MVNcall, allowing integration of multi allelic variants and complex events that weren't possible in phase 1.

   891k of the 1.37M sites missing from phase 1 were not identified by any phase 3 variant caller. These 891k SNPs have relatively high Ts/Tv ratio (1.84), which means these were likely missed in phase 3 because they are very rare, not because they are wrong; the increase in sample number in phase 3 made it harder to detect very rare events especially if the extra 1400 samples in phase 3 did not carry the alternative allele.

   481k of these SNPs were initially called in phase 3. 340k of them failed our initial SVM filter so were not included in our final merged variant set. 57k overlapped with larger variant events so were not accurately called. 84k sites did not make it into our final set of genotypes due to losses in our pipeline. Some of these sites will be false positives but we have no strong evidence as to which of these sites are wrong and which were lost for other reasons.

4. The reference genomes used for our alignments are different. Phase 1 alignments were aligned to the standard GRCh37 primary reference including unplaced contigs. In phase 3 we added EBV and a decoy set to the reference to reduce mismapping. This will have reduced our false positive variant calling as it will have reduced mismapping leading to false SNP calls. We cannot quantify this effect.

We have made no attempt to eludcidate why our SV and indel numbers changed. Since the release of phase 1 data, the algorithms to detect and validate indels and SVs have improved dramatically. By and large, we assume the indels and SVs in phase 1 that are missing from phase 3 are false positive in phase 1.

Project approach

If we can form clusters from a database of a sample of the world population by differentiating ethnicities ☾ via dataset 1000 genomes)

And that we project our sick patient database onto this sample (from PCAWG where the SNVs are represented)

And if we choose a panel of patients with mixed ethnicity (mestizo), it is potentially possible to identify within a pair of chromosomes, the paternal homologous chromosome as well as its maternal counterpart ☾ once this operation is done, it is

possible by the position of allels and snps to identify somatic mutations. and thus see where SNVs will tend to occur between the parent or mother chromosome.

DATABASES

1000 genomes dataset which includes the SNPs and their variations on the 23 pairs of chromosomes (healthy patient) HG19 VCF

	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00096	...	NA21128	NA21129	NA21130	NA21133	NA21135
0	22	16050075	.	A	G	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=8012;EA...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
1	22	16050115	.	G	A	100	PASS	AC=32;AF=0.00638978;AN=5008;NS=2504;DP=11468;E...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
2	22	16050213	.	C	T	100	PASS	AC=38;AF=0.00758786;AN=5008;NS=2504;DP=15092;E...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
3	22	16050319	.	C	T	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=22609;E...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
4	22	16050527	.	C	A	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=23591;E...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1931	22	51195015	.	G	C	100	PASS	AC=3;AF=0.000599042;AN=5008;NS=2504;DP=17466;E...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
1932	22	51195023	.	C	T	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504;DP=17120;E...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
1933	22	51195033	.	G	C	100	PASS	AC=32;AF=0.00638978;AN=5008;NS=2504;DP=17001;E...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
1934	22	51195041	.	A	G,T	100	PASS	AC=130,5;AF=0.0259585,0.000998403;AN=5008;NS=2...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0
1935	22	51195055	.	A	T	100	PASS	AC=4;AF=0.000798722;AN=5008;NS=2504;DP=16919;E...	GT	0|0	...	0|0	0|0	0|0	0|0	0|0

1936 rows × 2513 columns

We will work only on the exons → decreases the dataset + of interest

Dataset 1000 genomes (**SAMPLES .CSV**) → ethnies

```
Sample name     Sex     Biosample ID    Population code Population name Superpopulation code    Superpopulation name    Population elastic ID   Data collections
HG00315 female  SAME124335      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00327 female  SAME123791      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00334 female  SAME123981      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00339 female  SAME123991      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00341 male    SAME124922      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00346 female  SAME125264      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00353 female  SAME125120      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00358 male    SAME125126      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00360 male    SAME124539      FIN     Finnish,Finnish EUR     European Ancestry,West Eurasia (SGDP)   FIN,FinnishSGDP 1000 Genomes on GRCh38,Simons Genome Diversity Project,1000 Genomes 30x on
GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00365 female  SAME124542      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00372 male    SAME124748      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00377 female  SAME124752      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00384 female  SAME123034      FIN     Finnish EUR     European Ancestry       FIN     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00404 female  SAME123158      CHS     Southern Han Chinese    EAS     East Asian Ancestry     CHS     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1
release
HG00409 male    SAME1840026     CHS     Southern Han Chinese    EAS     East Asian Ancestry     CHS     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release
HG00411 male    SAME1839726     CHS     Southern Han Chinese    EAS     East Asian Ancestry     CHS
HG00106 female  SAME123948      GBR     British EUR     European Ancestry       GBR     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00113 male    SAME125342      GBR     British EUR     European Ancestry       GBR     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release
HG00118 female  SAME125347      GBR     British EUR     European Ancestry       GBR     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
HG00120 female  SAME122874      GBR     British EUR     European Ancestry       GBR     1000 Genomes on GRCh38,1000 Genomes 30x on GRCh38,1000 Genomes phase 3 release,1000 Genomes phase 1 release,Geuvadis
```

Allows us to classify ethnicities by overpopulation

Alleles.FILES --< positions of SNPs in chromosome pairs

Will allow us to work only on the desired positions

	position	a0	a1
0	16050115	3	1
1	16050213	2	4
2	16050607	3	1
3	16050783	1	3
4	16050840	2	3
...	...	...	...
384308	51241102	4	2
384309	51241285	4	3
384310	51241386	2	3
384311	51244163	1	3
384312	51244237	2	4

384313 rows × 3 columns

PCAWG → (sick patient) genome of diseased tissues where there are SNVs

File in hyperion → rar includes patients(and tt chromosomes

Planning

Step 1: PCA of SNPs from the 1000 Genome project (SNPs on exons that are present in PCAWG(is the buca patients etc …)) and projections of SNPs from PCAWG onto the principal components calculated on the 1000 Genome project. -> Allows identification of mixed ancestry samples in PCAWG.

Step 2: For 1 mixed ancestry sample, phase the SNPs to an ancestry by blocks of 500,000 bases (you may have to play with the window size later).

Step 3: For this sample, phase the SNVs to the SNPs and thus to the ancestry.

Step 4: Generate graphs or statistics on the somatic mutation rate on each parental copy for this sample.