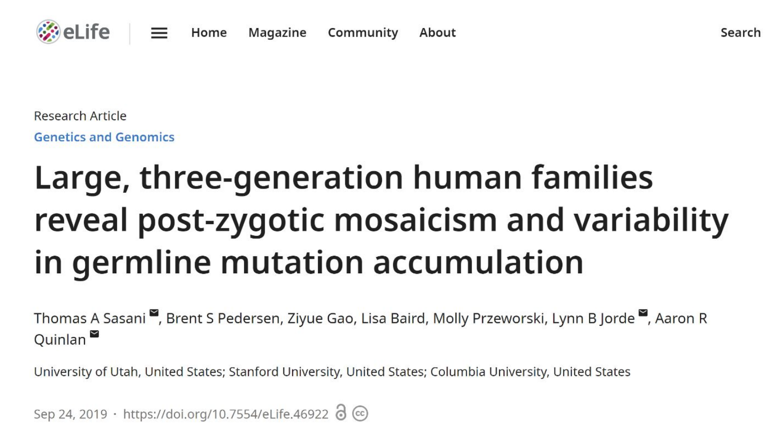
Looking at Genomic Data in R

We are all genetic mutants!

A baby will contain between 50 and 100 mutations that are not found in either parent

(number of mutations increases with the age of the father)



Research Article

Genetics and Genomics

Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation

Thomas A Sasani, Brent S Pedersen, Ziyue Gao, Lisa Baird, Molly Przeworski, Lynn B Jorde, Aaron R Quinlan

University of Utah, United States; Stanford University, United States; Columbia University, United States

Sep 24, 2019 · https://doi.org/10.7554/eLife.46922

Words you need to know

Reference allele: allele in reference genome
(most common; also called ancestral)

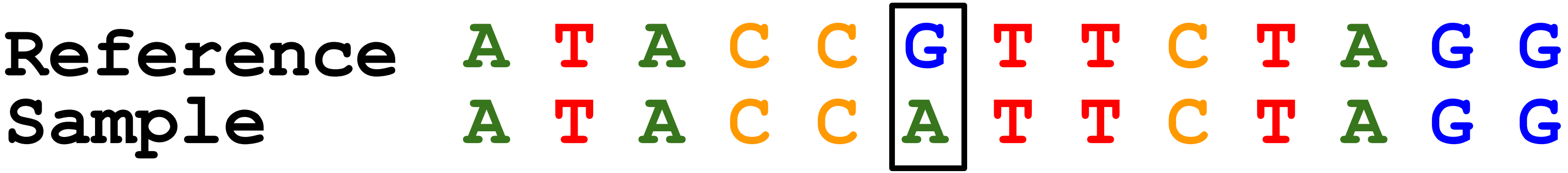
Alternate allele: mutation

Somatic mutation: mutation in DNA of body cells

Germline mutation: mutation in DNA of gametes

Types of Genomic Variation

Single Nucleotide Polymorphism (SNP)



```
Reference   A T A C C G T T C T A G G
Sample      A T A C C A T T C T A G G
```

SNP: alternate allele has MAF >1% in population

SNV: alternate allele has MAF <1% in population

Types of Genomic Variation

Insertion/Deletion (Indel)

Reference	A	T	A	C	C	G	T	T	-	-	C	T	A	G	G
Sample	A	T	A	C	C	G	T	T	G	G	C	T	A	G	G

Reference	T	T	G	C	A	A	A	G	G	A	T	C	C
Sample	T	T	G	-	A	A	A	G	G	A	T	C	C

Types of Genomic Variation

Copy Number Variant (CNV)

Reference	T	T	G	T	-	-	-	-	-	-	A	A	A	G	G
Sample 1	T	T	G	T	G	T	-	-	-	-	A	A	A	G	G
Sample 2	T	T	G	T	G	T	G	T	G	T	A	A	A	G	G

Exercise: Genomic Data

Common types of genomic data files

FASTQ

FASTA

SAM/BAM

BED

VCF

Variant Call Format

Variant call format (VCF) file - header

```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

http://www.internationalgenome.org/wiki/Analysis/vcf4.0/

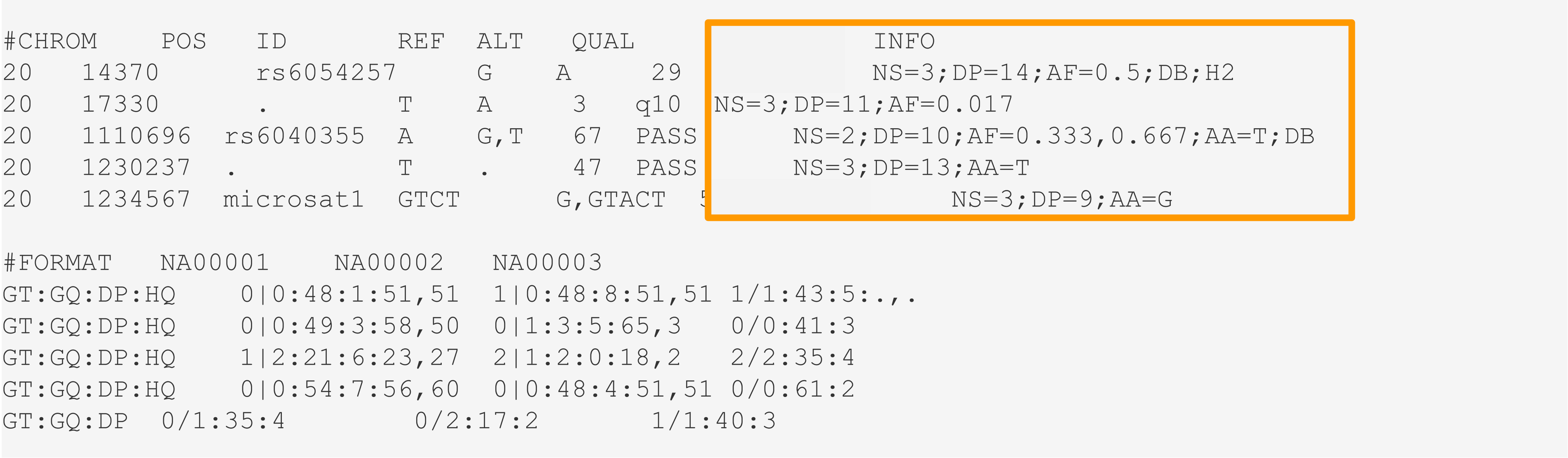
Variant call format (VCF) file – body of the file

```
#CHROM    POS   ID       REF  ALT   QUAL     FILTER    INFO
20   14370      rs6054257     G    A     29  PASS      NS=3;DP=14;AF=0.5;DB;H2
20   17330      .        T    A     3   q10  NS=3;DP=11;AF=0.017
20   1110696  rs6040355  A    G,T   67  PASS      NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20   1230237  .          T    .     47  PASS      NS=3;DP=13;AA=T
20   1234567  microsat1  GTCT      G,GTACT  50    PASS      NS=3;DP=9;AA=G

#FORMAT   NA00001    NA00002   NA00003
GT:GQ:DP:HQ    0|0:48:1:51,51  1|0:48:8:51,51 1/1:43:5:.,.
GT:GQ:DP:HQ    0|0:49:3:58,50  0|1:3:5:65,3   0/0:41:3
GT:GQ:DP:HQ    1|2:21:6:23,27  2|1:2:0:18,2   2/2:35:4
GT:GQ:DP:HQ    0|0:54:7:56,60  0|0:48:4:51,51 0/0:61:2
GT:GQ:DP  0/1:35:4        0/2:17:2       1/1:40:3
```

http://www.internationalgenome.org/wiki/Analysis/vcf4.0/

Variant call format (VCF) file – body of the file



```
#CHROM     POS    ID         REF   ALT    QUAL                    INFO
20   14370       rs6054257       G     A      29                  NS=3;DP=14;AF=0.5;DB;H2
20   17330       .          T     A      3   q10   NS=3;DP=11;AF=0.017
20   1110696   rs6040355  A     G,T    67   PASS        NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20   1230237   .          T     .      47   PASS        NS=3;DP=13;AA=T
20   1234567   microsat1  GTCT       G,GTACT   5                  NS=3;DP=9;AA=G

#FORMAT   NA00001     NA00002    NA00003
GT:GQ:DP:HQ     0|0:48:1:51,51  1|0:48:8:51,51 1/1:43:5:.,.
GT:GQ:DP:HQ     0|0:49:3:58,50  0|1:3:5:65,3   0/0:41:3
GT:GQ:DP:HQ     1|2:21:6:23,27  2|1:2:0:18,2   2/2:35:4
GT:GQ:DP:HQ     0|0:54:7:56,60  0|0:48:4:51,51 0/0:61:2
GT:GQ:DP  0/1:35:4        0/2:17:2       1/1:40:3
```

AA - ancestral allele

AC - allele count in genotypes for each ALT allele (same order as listed)

AF - allele frequency for each ALT allele (same order as listed; used for primary data)

DB - dbSNP membership

H2 - membership in hapmap2

http://www.internationalgenome.org/wiki/Analysis/vcf

Variant call format (VCF) file – body of the file

```
#CHROM    POS   ID       REF  ALT   QUAL     FILTER    INFO
20   14370      rs6054257     G    A     29  PASS      NS=3;DP=14;AF=0.5;DB;H2
20   17330      .        T    A     3   q10  NS=3;DP=11;AF=0.017
20   1110696  rs6040355  A    G,T   67  PASS      NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20   1230237  .          T    .     47  PASS      NS=3;DP=13;AA=T
20   1234567  microsat1  GTCT      G,GTACT  50    PASS      NS=3;DP=9;AA=G

#FORMAT   NA00001    NA00002   NA00003
GT:GQ:DP:HQ    0|0:48:1:51,51  1|0:48:8:51,51 1/1:43:5:.,.
GT:GQ:DP:HQ    0|0:49:3:58,50  0|1:3:5:65,3   0/0:41:3
GT:GQ:DP:HQ    1|2:21:6:23,27  2|1:2:0:18,2   2/2:35:4
GT:GQ:DP:HQ    0|0:54:7:56,60  0|0:48:4:51,51 0/0:61:2
GT:GQ:DP  0/1:35:4        0/2:17:2       1/1:40:3
```

GT - genotype
GQ - genotype quality
HQ - haplotype quality
DP - combined depth across samples, e.g. DP=154
NS - Number of samples with data

http://www.internationalgenome.org/wiki/Analysis/vcf

Variant call format (VCF) file – body of the file

```
#CHROM     POS    ID         REF   ALT    QUAL      FILTER     INFO
20   14370      rs6054257      G     A      29   PASS       NS=3;DP=14;AF=0.5;DB;H2
20   17330      .          T     A      3   q10  NS=3;DP=11;AF=0.017
20   1110696  rs6040355  A     G,T    67   PASS       NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20   1230237  .          T     .      47   PASS       NS=3;DP=13;AA=T
20   1234567  microsat1  GTCT        G,GTACT  50     PASS       NS=3;DP=9;AA=G

#FORMAT    NA00001     NA00002    NA00003
GT:GQ:DP:HQ     0|0:48:1:51,51   1|0:48:8:51,51  1/1:43:5:.,.
GT:GQ:DP:HQ     0|0:49:3:58,50   0|1:3:5:65,3    0/0:41:3
GT:GQ:DP:HQ     1|2:21:6:23,27   2|1:2:0:18,2    2/2:35:4
GT:GQ:DP:HQ     0|0:54:7:56,60   0|0:48:4:51,51  0/0:61:2
GT:GQ:DP  0/1:35:4        0/2:17:2        1/1:40:3
```

0|0: two copies of reference allele

1|0: one reference allele, one alternate allele

1|2: two different alternate alleles

1/0: one reference allele, one alternate allele, but genome *unphased*

http://www.internationalgenome.org/wiki/Analysis/vcf4.0/

Genomes make no sense without a guidebook

Annotation:
the process of describing the structure and function of the components of a genome

This helps us figure out what variants might be important for our research question.

Human Genome Variant Databases

1000 Genomes: https://www.internationalgenome.org/

dbSNP: https://www.ncbi.nlm.nih.gov/snp/

dbVar: https://www.ncbi.nlm.nih.gov/dbvar/

ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/

Exome Aggregation Consortium (ExAC): https://gnomad.broadinstitute.org/downloads#exac-variants

Genome Aggregation Database (gnomAD): https://gnomad.broadinstitute.org/

Genome Data Commons: https://gdc.cancer.gov/

Annotation Tools:

ANNOVAR: https://annovar.openbioinformatics.org/en/latest/

SnpEff: https://pcingola.github.io/SnpEff/

SIFT: https://sift.bii.a-star.edu.sg/

GATK VariantAnnotator

VariantAnnotation R Package:
https://bioconductor.org/packages/release/bioc/html/VariantAnnotation.html

Variant Annotation Integrator (UCSC):
https://genome.ucsc.edu/cgi-bin/hgVai

biomaRT:
https://bioconductor.org/packages/release/bioc/html/biomaRt.html

Exercise: Variant Calls