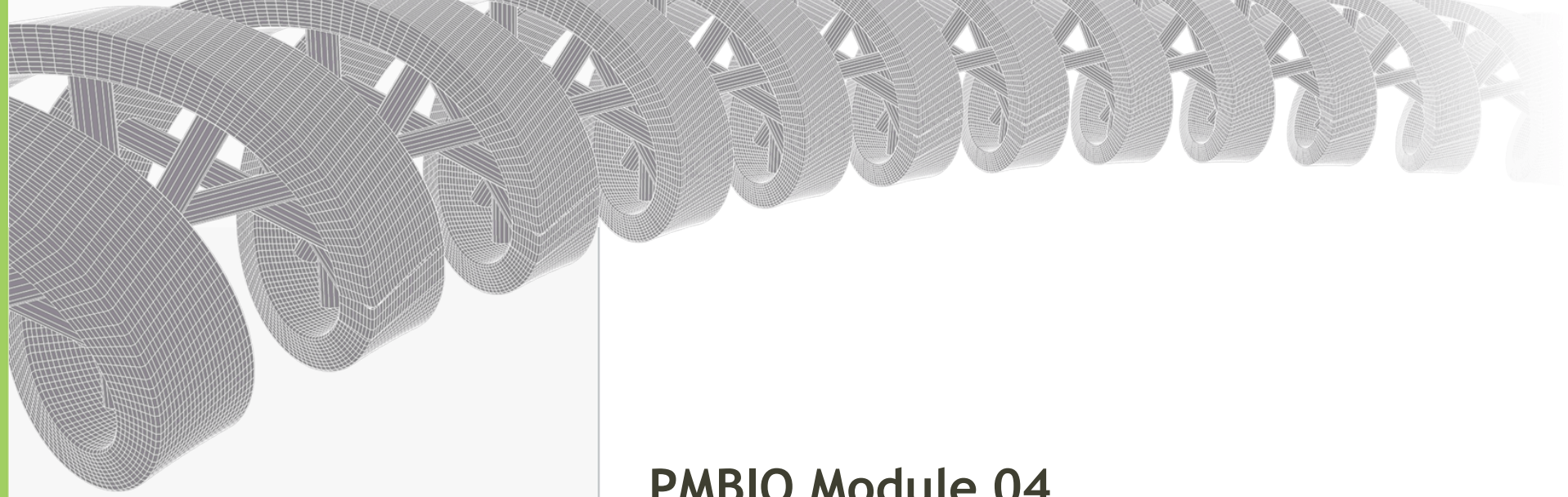


The Elizabeth H. and James S. McDonnell III
McDonnell Genome Institute
at Washington University

Washington University in St. Louis
School of Medicine

# PMBIO Module 04

## **Germline.** Germline WGS and Exome Variant Analysis

Malachi Griffith, Obi Griffith, Zachary Skidmore, Huiming Xia
Introduction to bioinformatics for DNA and RNA sequence analysis (IBDR01)
29 October – 2 November, 2018
Glasgow

# Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

## You are free to:


**Share** — copy and redistribute the material in any medium or format

**Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

## Under the following terms:

**Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

**ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

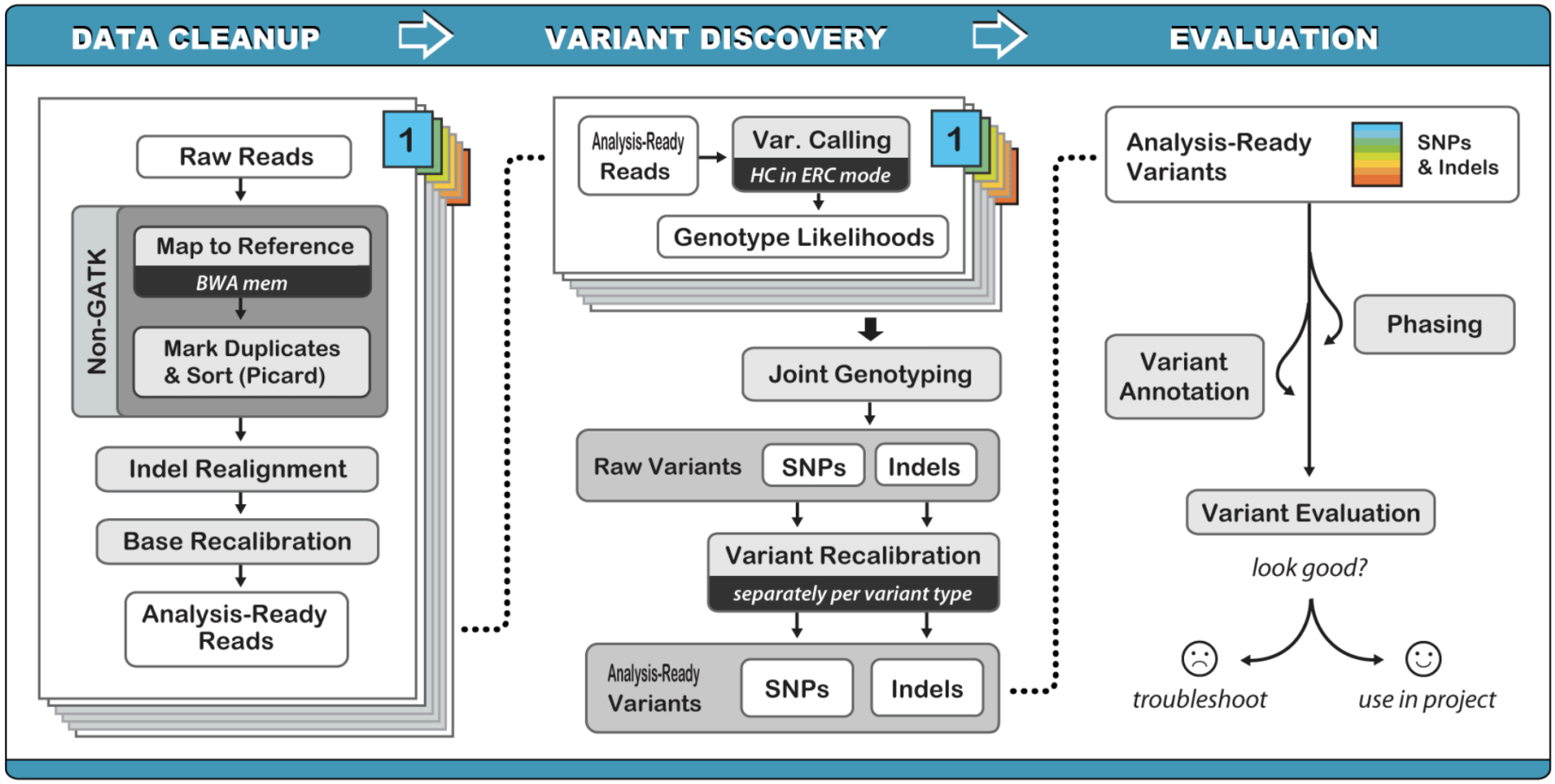
**No additional restrictions** — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

# Learning objectives of module 04: Germline

- **Key concepts**: Variation, germline mutation, polymorphism, SNPs, variant databases, germline variant calling

- Use of the GATK tool kit for variant calling
- Perform basic germline variant filtering
- Compare single sample germline variant calling to cohort based calling

# An example germline variant workflow

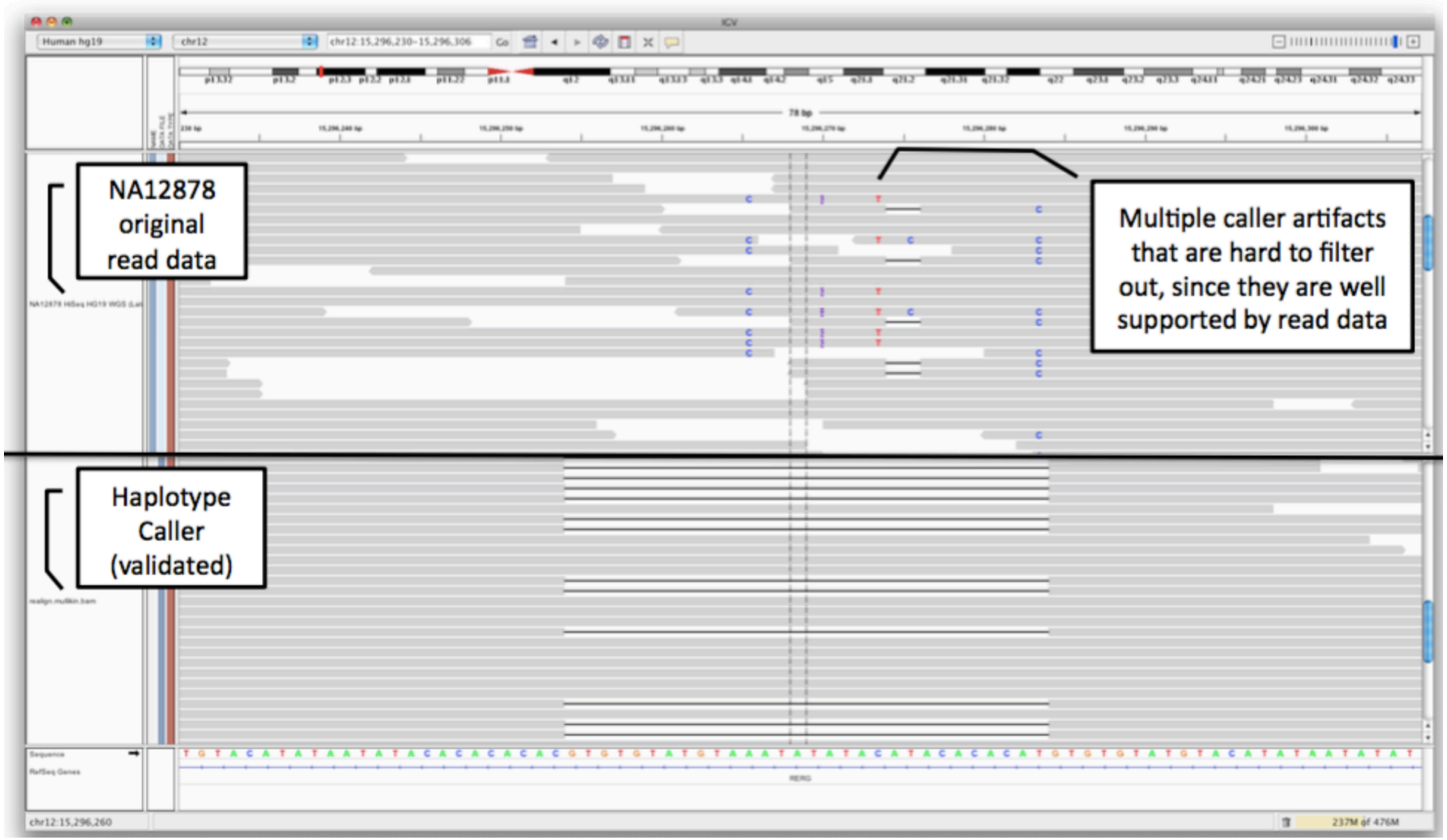



Slide courtesy of Karyn Meltz Steinberg (Washington University)

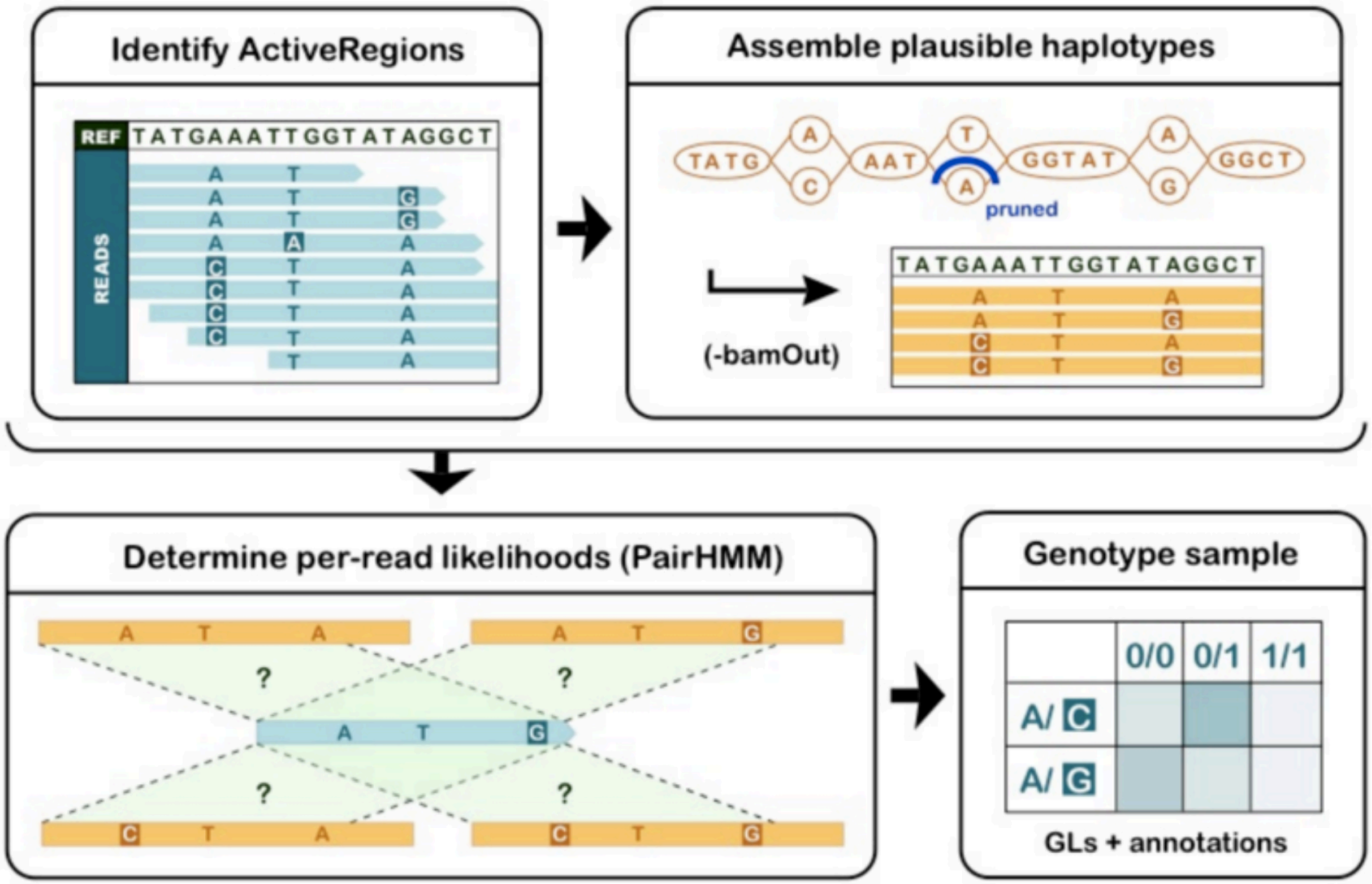
# Different methods

- Call SNVs and Indels separately by considering each variant locus
  - Very fast
  - Assumes bases are independent
- Call SNVs and indels simultaneously via Bayesian genotype likelihood model
  - More computationally intensive
  - GATK UnifiedGenotyper
- Call SNVs, indels and SVs simultaneously by performing a local de novo assembly
  - More computationally intensive
  - More accurate—gets rid of many false positives especially indels • GATK HaplotypeCaller

# Why all the steps?



# Haplotype caller illustrated



# GATK recommended filters

- SNPs
  - QD < 2.0 (variant quality/depth of non-ref samples)
  - MQ < 40.0 (Mapping quality)
  - FS > 60.0 (Phred score Fisher's test pvalue for strand bias)
  - SOR > 3.0 (Strand odds ratio, aims to evaluate whether there is strand bias in the data—updated form of FET)
  - MQRankSum < -12.5 (mapping quality of reference reads vs alt reads)
  - ReadPosRankSum < -8.0 (distance of alt reads from end of the read)
- Indels
  - QD<2.0
  - ReadPosRankSum < -20.0 • InbreedingCoeff < -0.8
  - FS > 200.0
  - SOR > 10.0

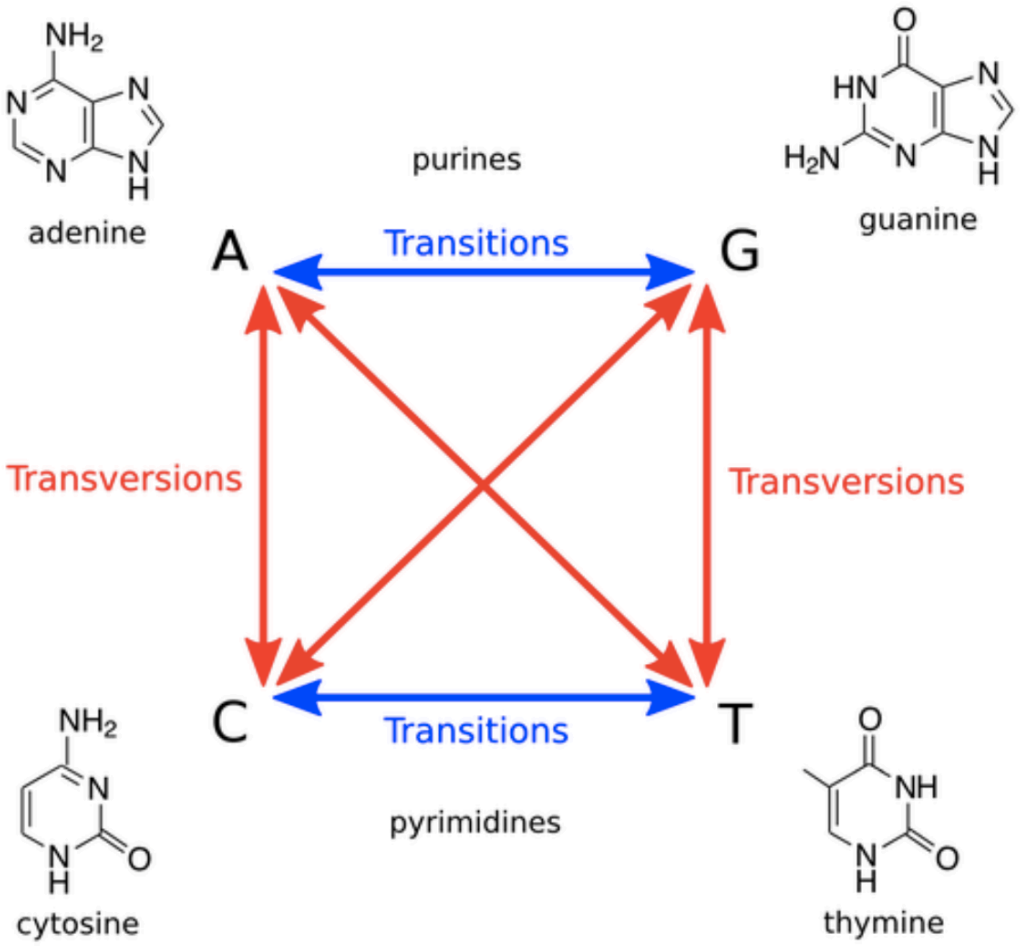
# Variant Quality Score Recalibration (VSQR)

- VariantRecalibrator
  - Gaussian mixture model by looking at the annotation values over a high quality subset of input call set and then evaluate all variants
- ApplyRecalibration
  - Apply model parameters to each variant producing a recalibrated VCF file

- Run these separately for SNPs and indels
- Need to have some resource files to train the calibrator
  - HapMap
  - 1000G
  - dbSNP

# Variant evaluation

# Transition/Transversion ratio (Ti/Tv)



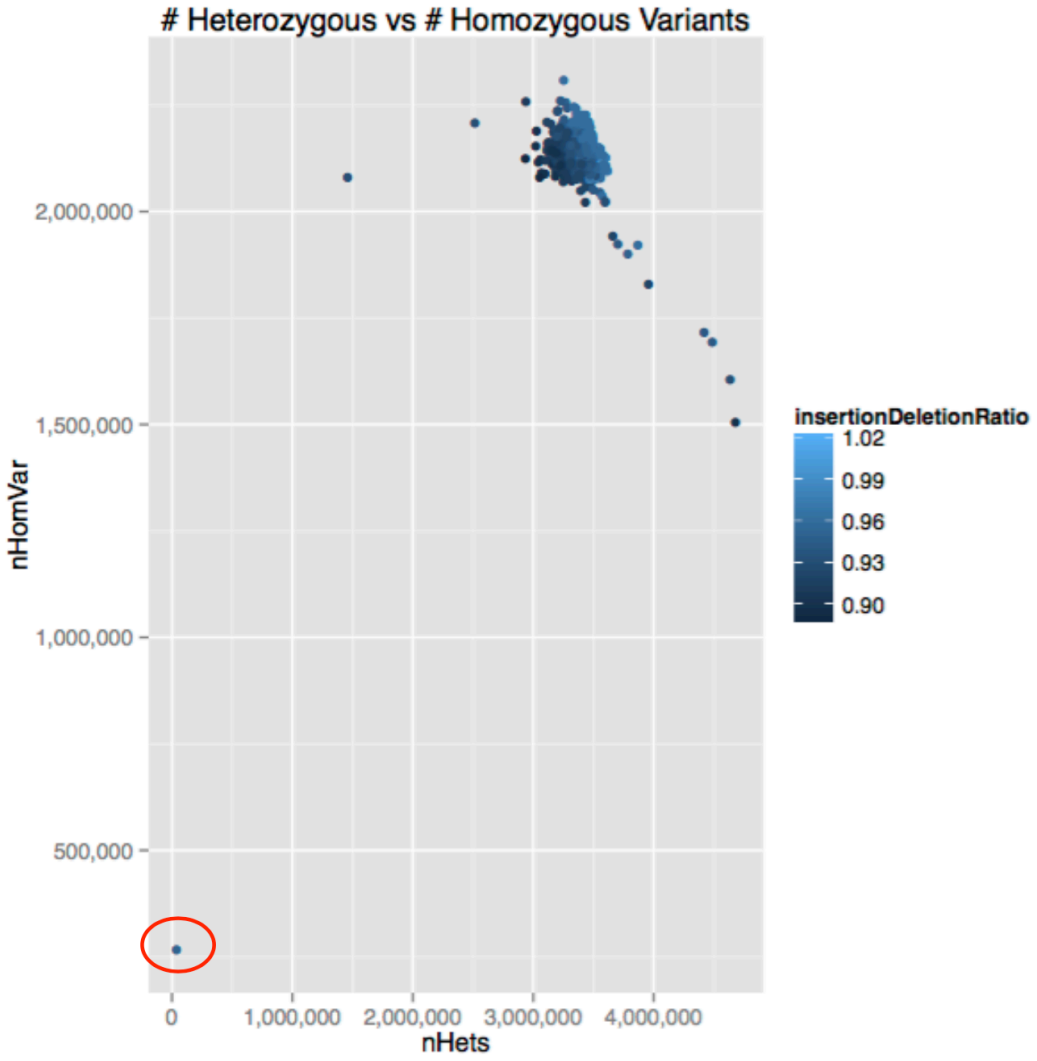

Random = 0.5

WGS = 2.0-2.1

Exome = 3-3.5

Watch for major deviation from typical ratio

# # heterozygous and homozygous variant




Watch for outliers …

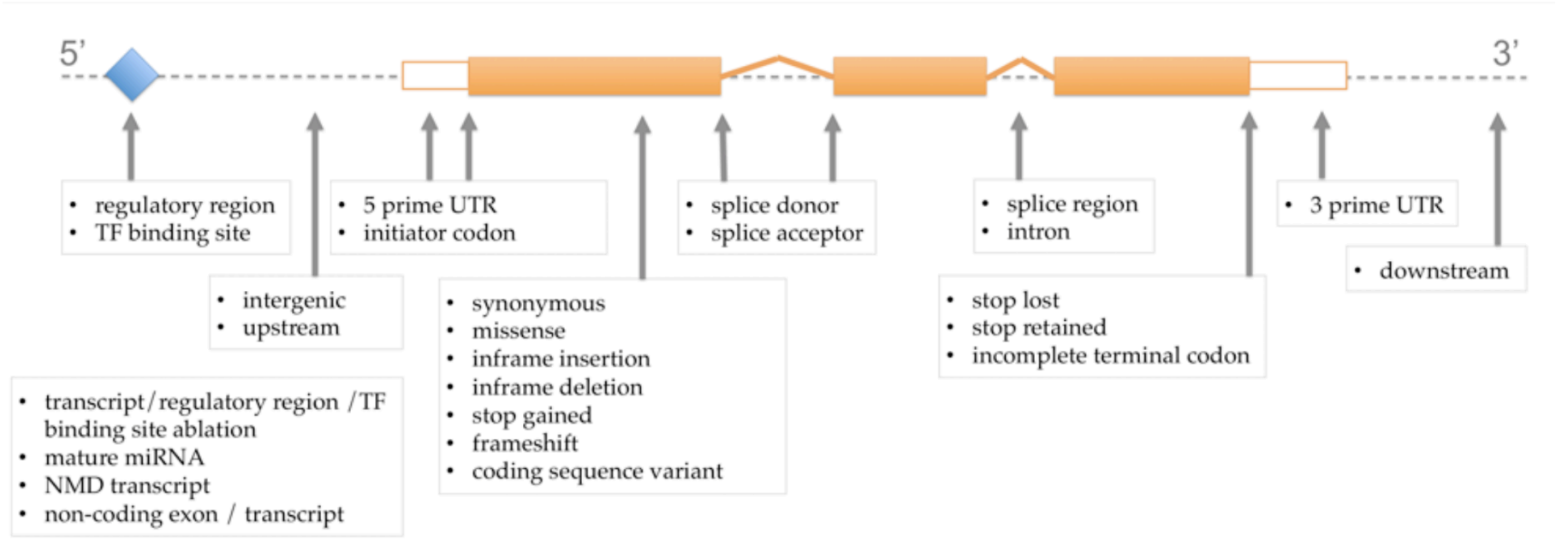
# Other things to evaluate

- Sample swaps? Unexpected duplicates?
- Sex check
- Novel vs Known sites
- dbSNP/1KG concordance (should be >98%) • Number of singletons per sample
- excess singletons may indicate a bad sample
- PCA
  - Is there significant population stratification?
  - Do you expect this?
  - What could be causing? Batch effects?

# Germline variant annotation approaches/tool

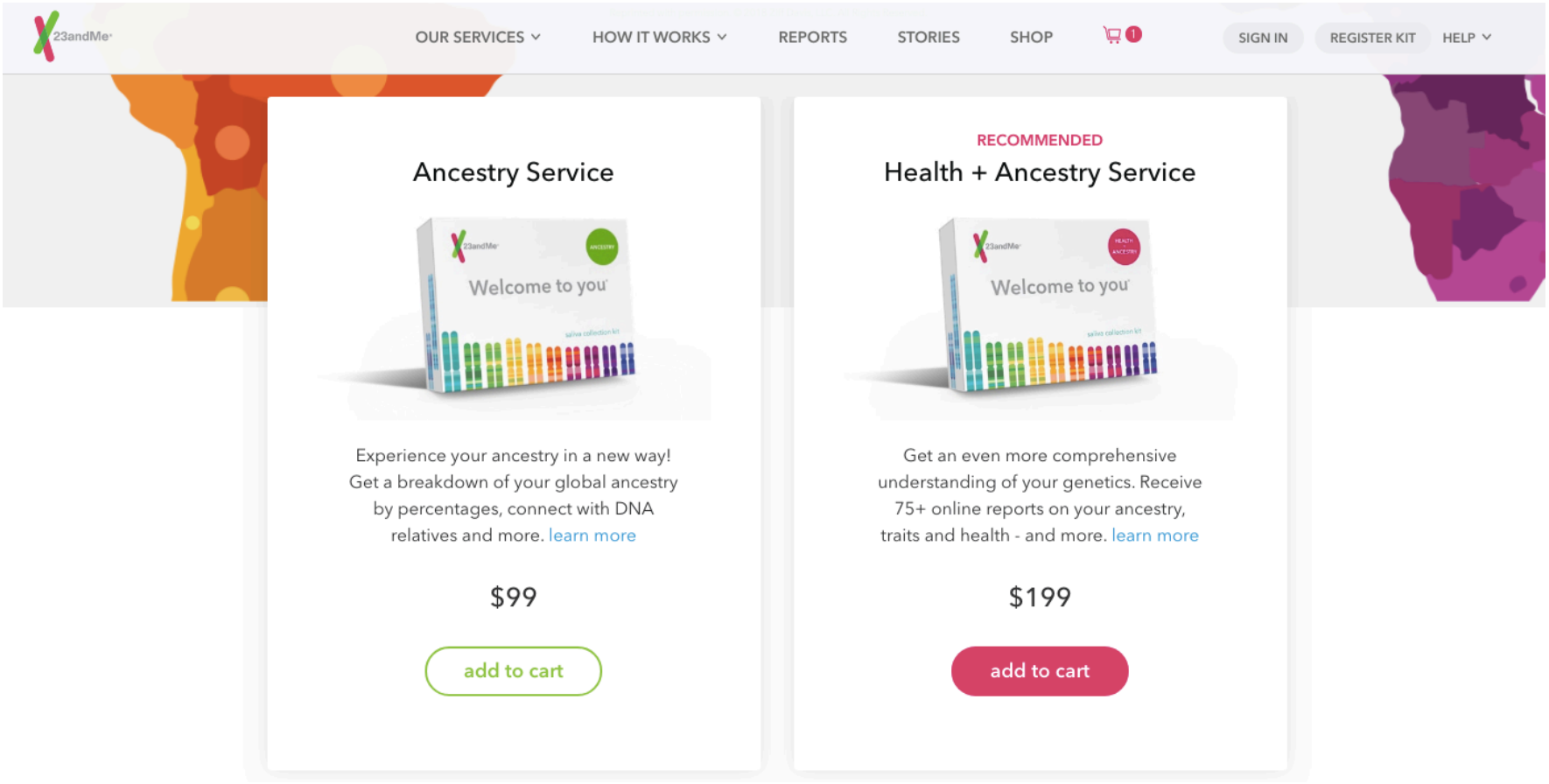
- VEP
  - Variant Effect Predictor
- SIFT
  - Sorting Intolerant From Tolerant
  - SIFTpredictswhetheranaminoacidsubstitutionaffects
  - protein function
  - Given a protein sequence, SIFT chooses related proteins and obtains an alignment of these proteins with the query. Based on the amino acids appearing at each position in the alignment, SIFT calculates the probability that an amino acid at a position is tolerated conditional on the most frequent amino acid being tolerated. If this normalized value is less than a cutoff, the substitution is predicted to be deleterious
- PolyPhen
  - Polymorphism Phenotyping
  - Predicts possible impact of amino acid substitution on structure and function of human protein using physical and comparative considerations
- CADD
  - CombinedAnnotationDependentDepletion
  - Integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations
- GEMINI
  - Helps to aggregate and perform complex queries involving many variant annotations including those above

# Functional categorization relative to transcript annotations



# Personal anecdote on personal (direct to consumer) genomics

# What is consumer genomics?




- Direct-to-consumer genetic testing refers to genetic tests that are marketed directly to consumers. Also referred to as “at-home” genetic testing.
- Provide access to a person’s genetic information without *necessarily* involving a doctor or insurance company in the process.
  - Our “ask your doctor …” culture blurs the lines between DTC and medically indicated products and services
- Allows consumers to learn about ancestry, DNA relatives, genetic traits, health predispositions, family planning, etc, “on their own”.

# Current Consumer Genomics options

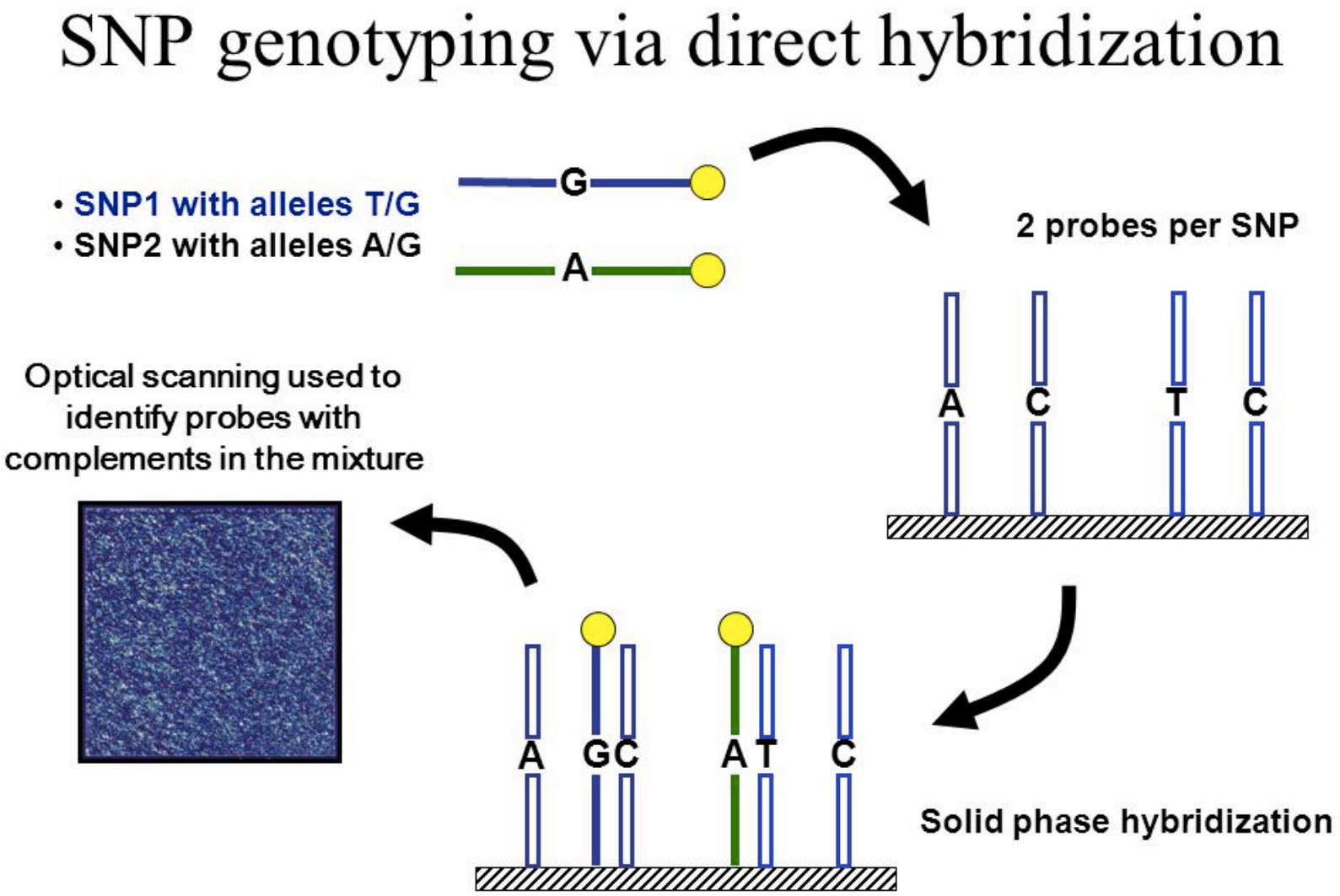
- Primarily genotyping but increasingly also some exome, whole genome, transcriptome and other options
  - Ancestry
  - Genotype-phenotype associations
  - Health risks
  - Early disease detection - intervention
  - Undiagnosed disease
  - Family planning
  - Precision medicine

- Representative companies:
  - 23andMe (DTC)
  - Ancestry.com (DTC)
  - Gene By Gene (DTC)
  - FullGenomes (DTC)
  - MyMedLab[?]
  - Color Genomics*
  - Genos*
  - Veritas*
  - HLI*
  - Illumina*
  - Counsyl*
  - Foundation Medicine*
  - Many more…

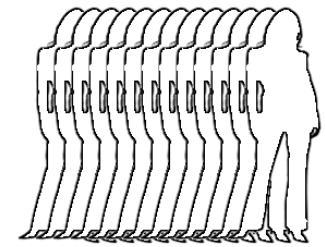
* Physician referral required

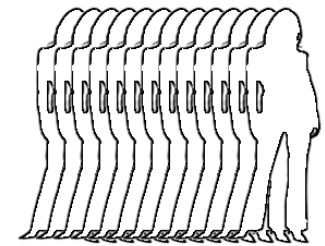
[?] Unclear if referral needed

# SNP array genotyping identifies likely genotype by hybridization



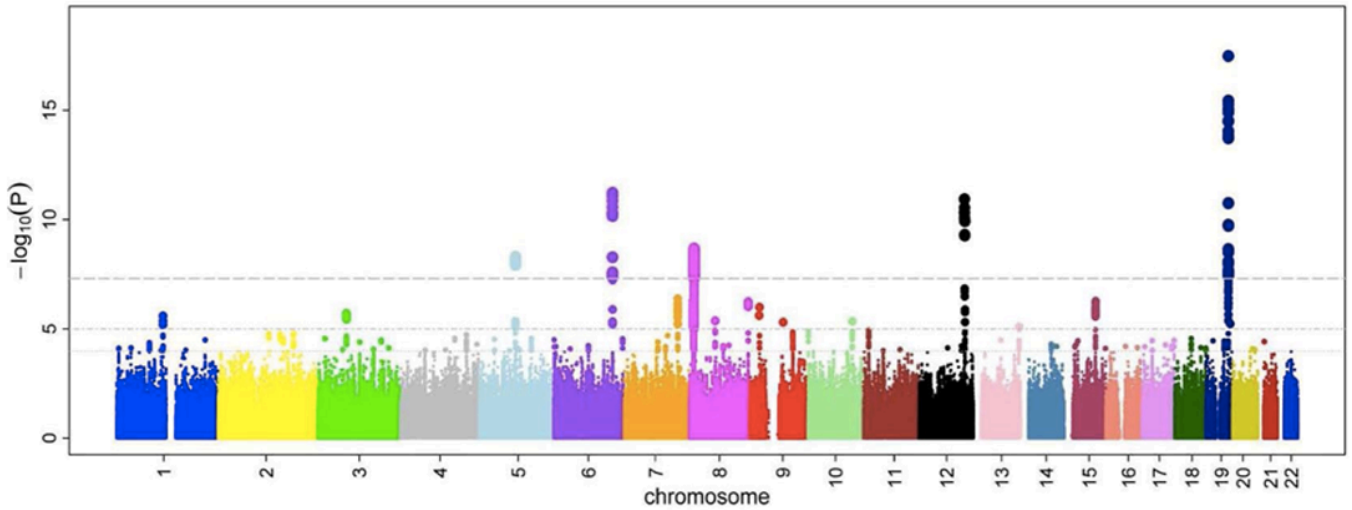
# Inferred genotypes are intersected with known genotype-phenotype associations (GWAS studies)




| *SNP1* | *SNP2* | *SNP…* |
|---|---|---|
| **Cases**<br>Count of G:<br>2104 of 4000<br><br>Frequency of G:<br>52.6% | **Cases**<br>Count of G:<br>1648 of 4000<br><br>Frequency of G:<br>41.2% | *Repeat for all SNPs* |
| **Controls**<br>Count of G:<br>2676 of 6000<br><br>Frequency of G:<br>44.6% | **Controls**<br>Count of G:<br>2532 of 6000<br><br>Frequency of G:<br>42.2% | |
| **P-value:**<br>$5.0 \cdot 10^{-15}$ | **P-value:**<br>0.33 | |

- The frequencies of specific alleles at SNP positions are statistically compared between cases and controls for a particular disease/trait
- Many SNPs are scanned across the whole genome and significant loci identified
- Thousands of such studies have now been completed

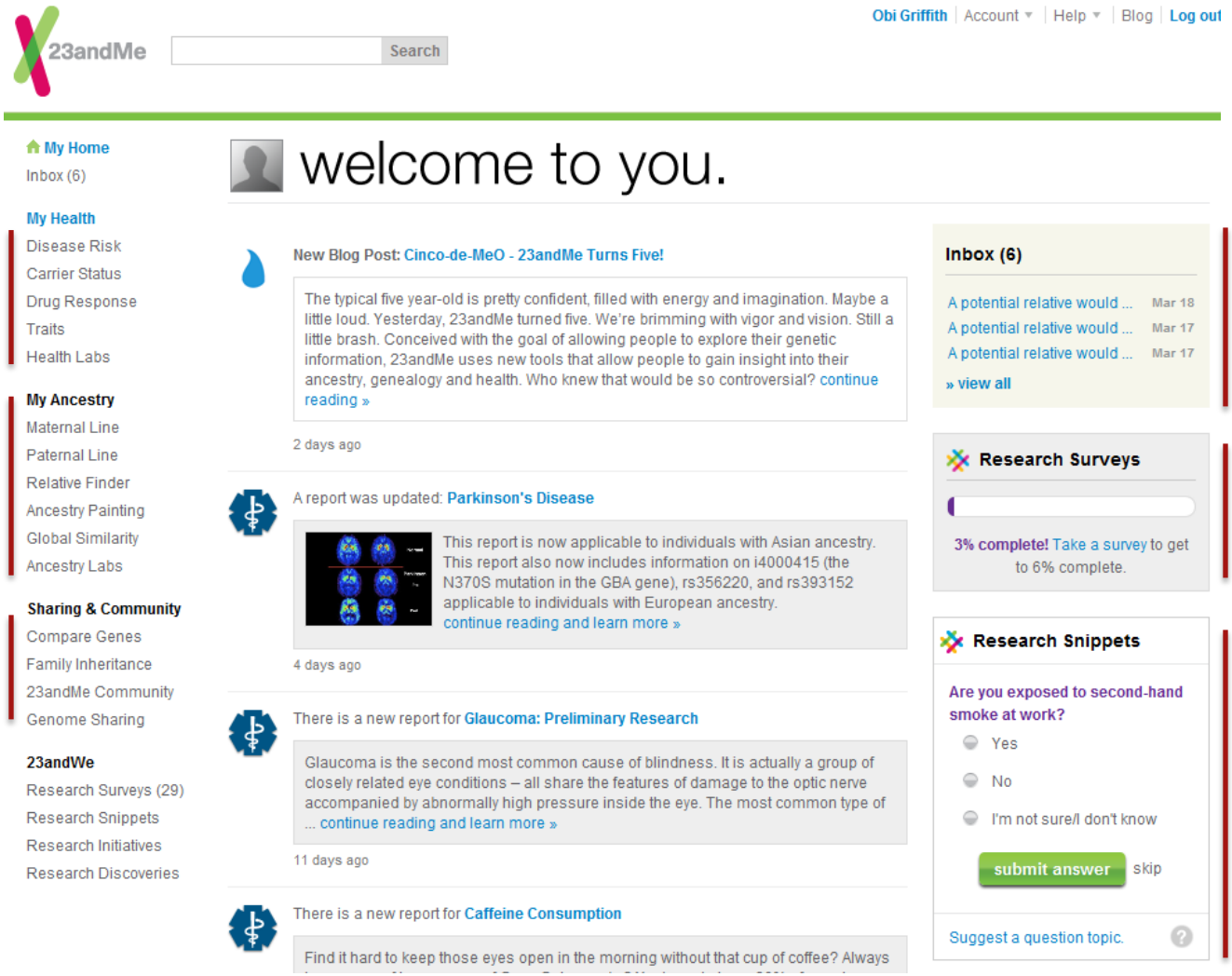


# 23andMe: Welcome to me

Health risks
Traits, etc

Ancestry

Sharing
Comparing



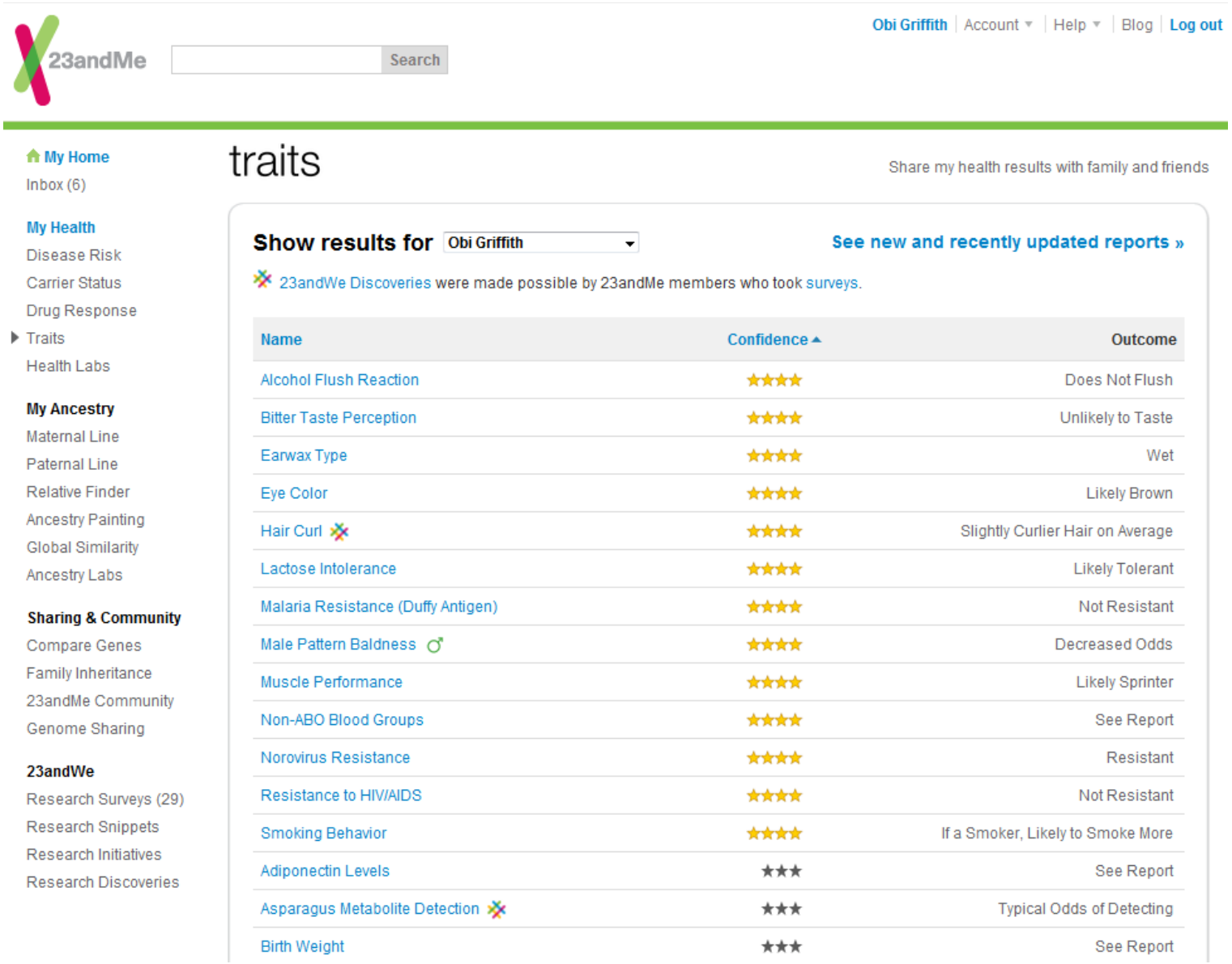

Potential relatives

Surveys

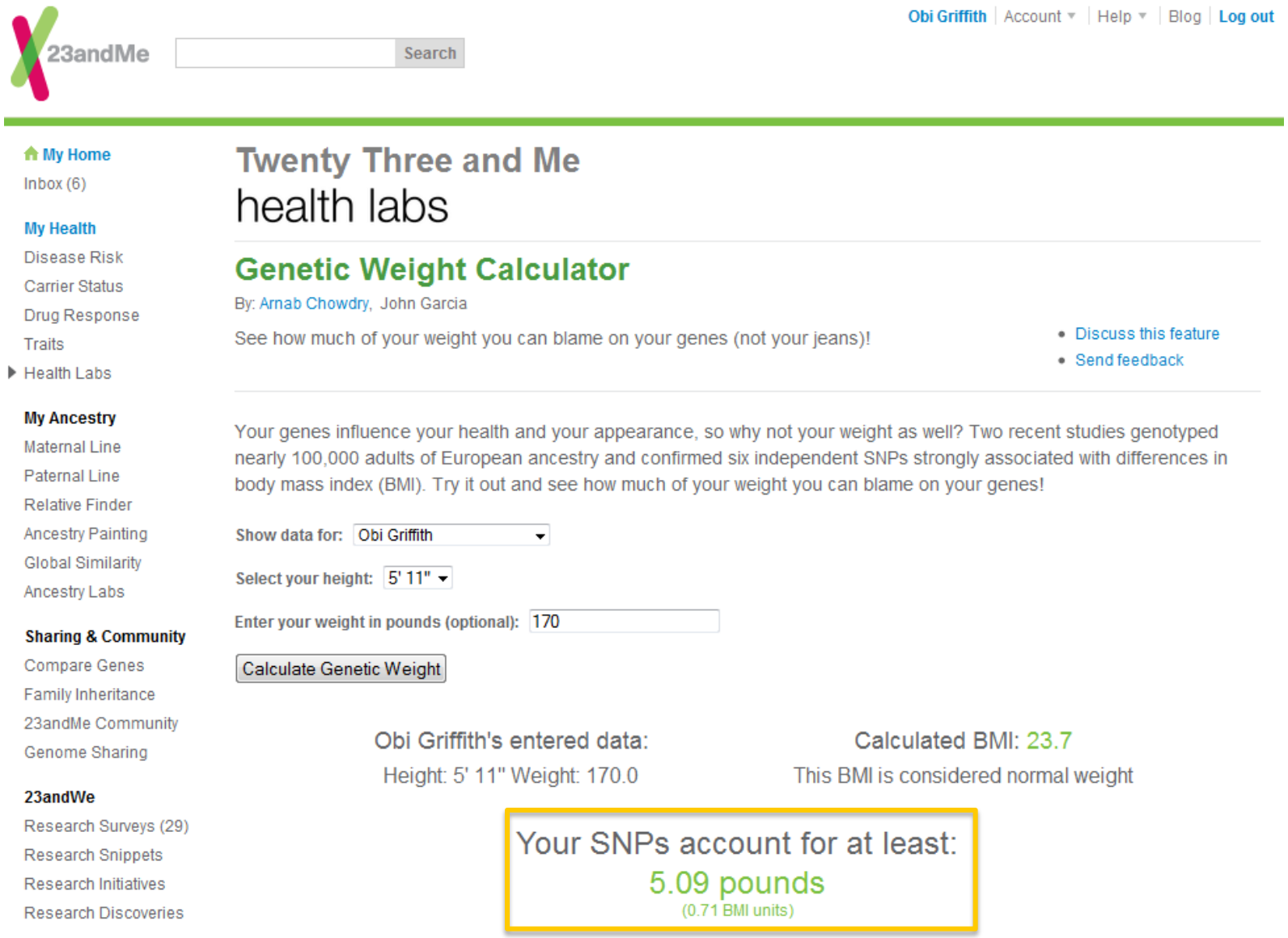
Snippets

Why did I do 23andMe? To compare similarities and differences with my fraternal (dizygotic) twin

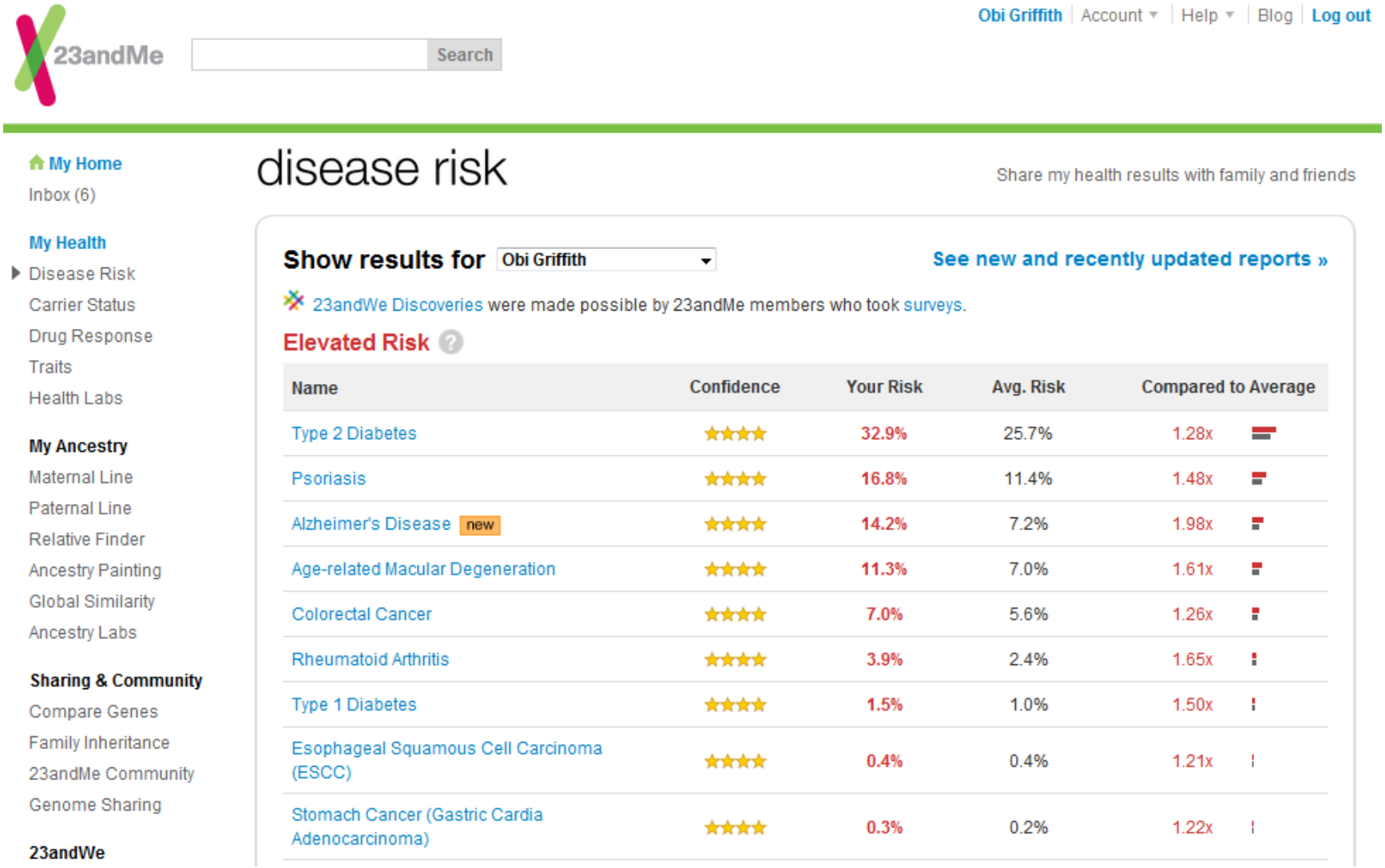
# Traits: A “likely sprinter” with wet earwax



# “Health Labs”: Extra five pounds not my fault

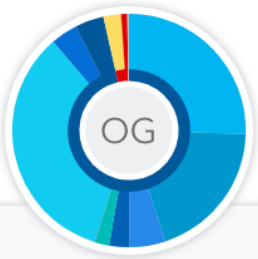


# Disease risks - 33% chance of diabetes! Oh wait that's just 1.3x average

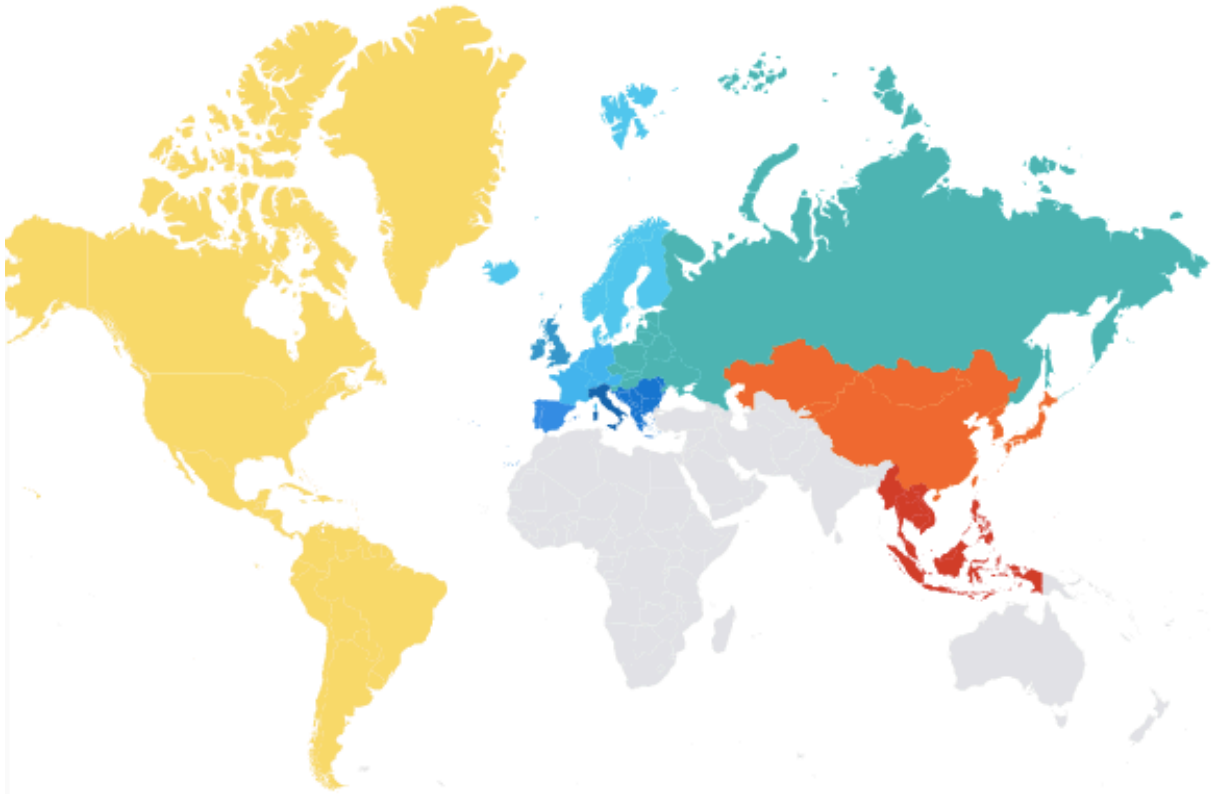



Environmental factors (not considered by 23andMe) likely outweigh my increased risk due to genetics

# 23andMe says - 96.5% European with 3.3% (East Asian & Native American)




| Obi Griffith | 100% |
|---|---|
| **European** | **96.5%** |
| French & German | 25.5% |
| British & Irish<br>United Kingdom | 19.5% |
| Iberian | 4.9% |
| Italian | 2.8% |
| Eastern European | 1.9% |
| Ashkenazi Jewish | 0.1% |
| Broadly Northwestern European | 34.0% |
| Broadly Southern European | 3.7% |
| Broadly European | 4.0% |
| **East Asian & Native American** | **3.3%** |
| Native American | 2.4% |
| Broadly East Asian | 0.1% |
| Broadly East Asian & Native American | 0.8% |
| **Unassigned** | **0.2%** |



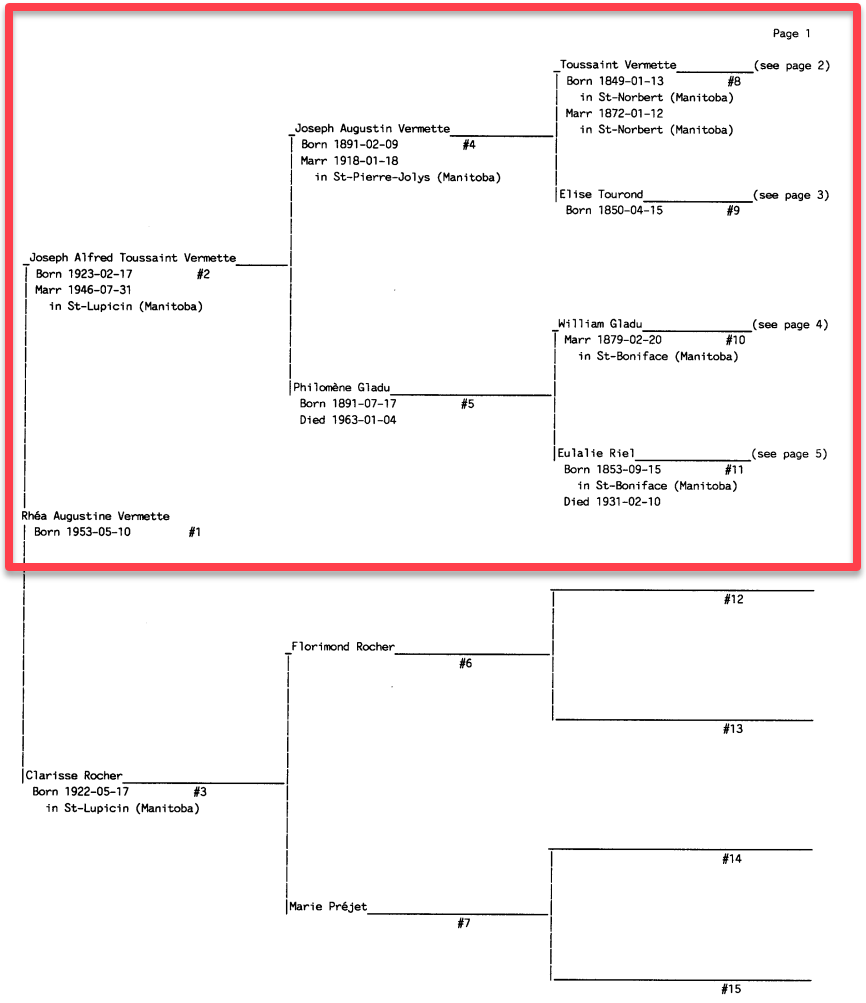
Where did this Asian or Native American component come from?

# To the records office…

Grandfather

Mother

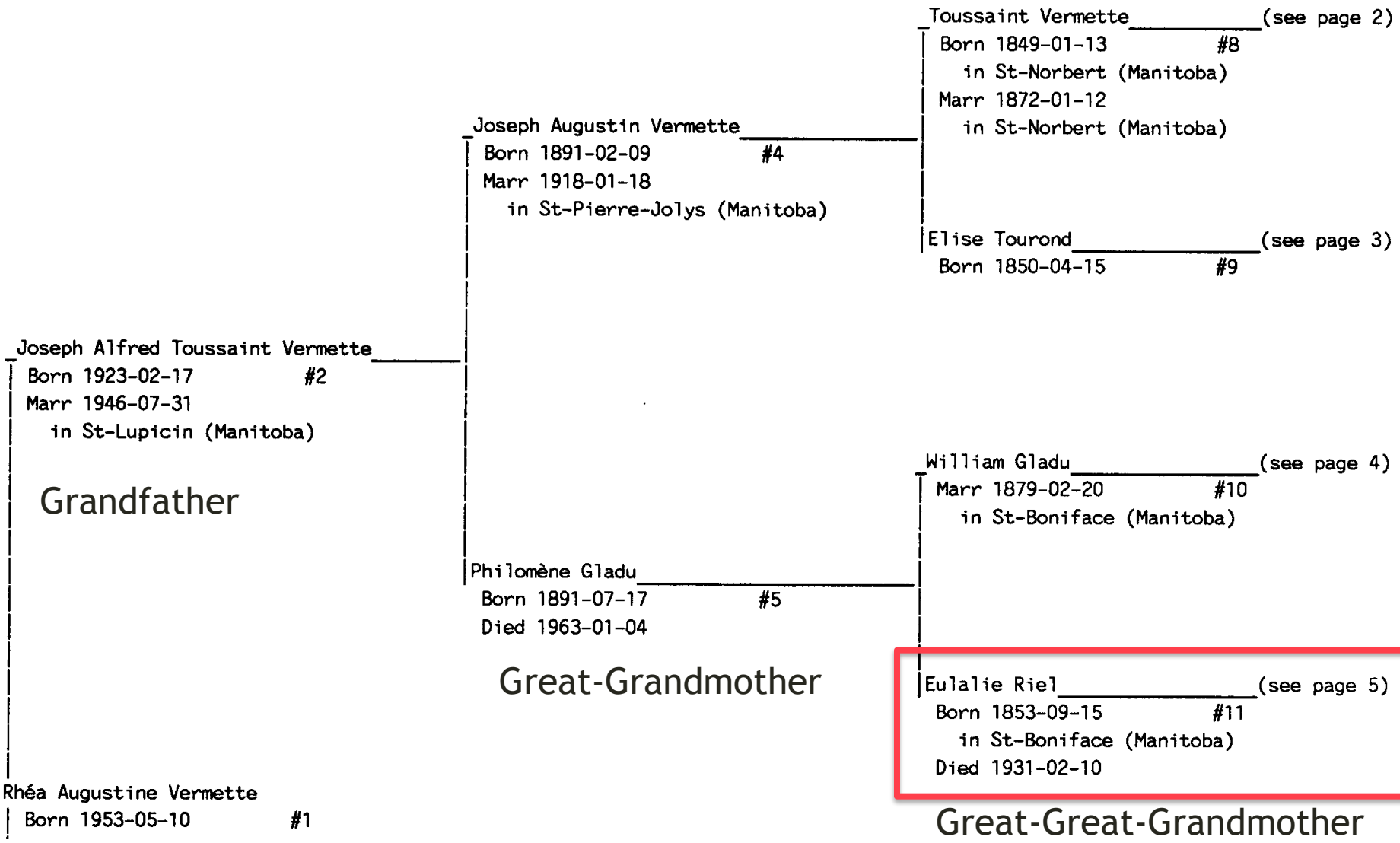
Grandmother




Starting from my mother's name and birthdate, they were able to provide 71 pages of ancestry tracing as far back as 1574 France

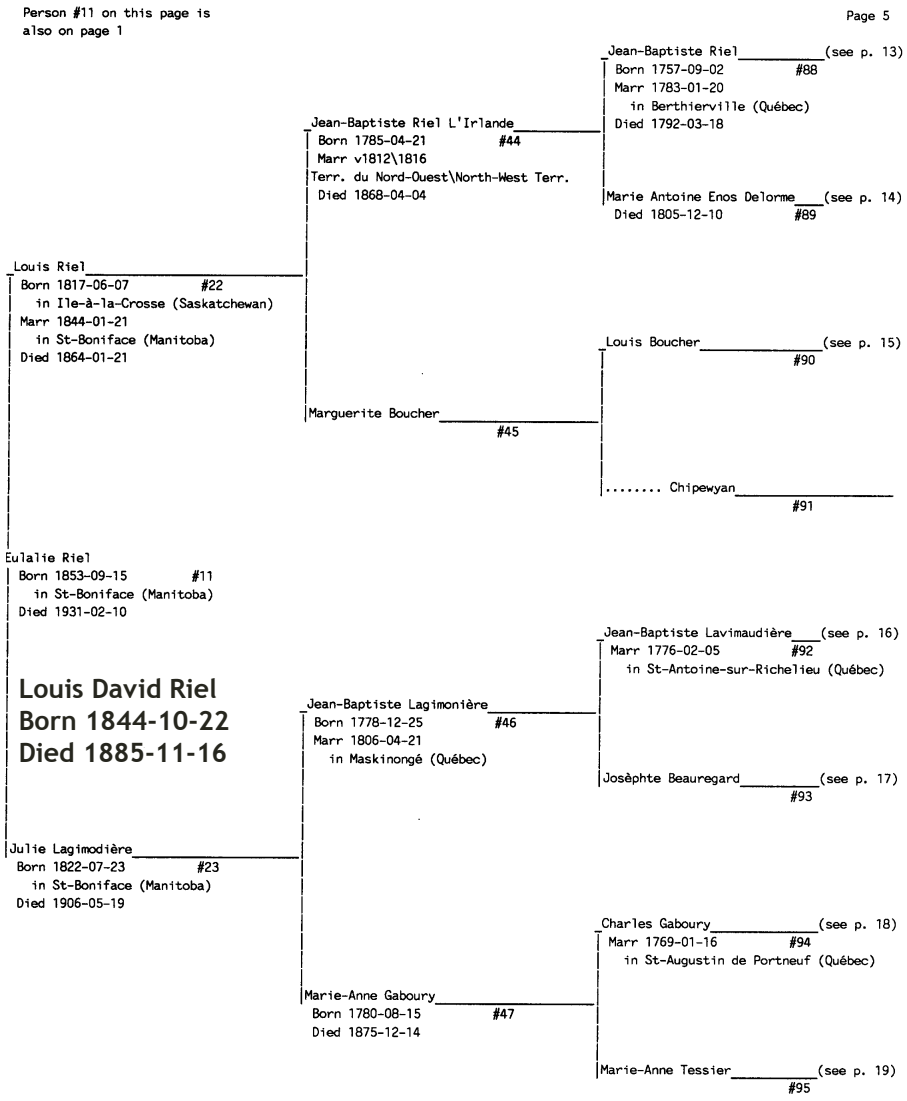
The Société historique de Saint-Boniface - an archives and research center dedicated to preserving, studying, disseminating and promoting the francophone and Métis history of Manitoba and Western Canada.

# To the records office...



# An historical aside…

Great-Great-Grandmother



# An historical aside…

Great-Great-Grandmother

Person #11 on this page is also on page 1

Louis Riel
Born 1817-06-07 #22
in Ile-à-la-Crosse (Saskatchewan)
Marr 1844-01-21
in St-Boniface (Manitoba)
Died 1864-01-21

Eulalie Riel
Born 1853-09-15 #11
in St-Boniface (Manitoba)
Died 1931-02-10

**Louis David Riel**
**Born 1844-10-22**
**Died 1885-11-16**

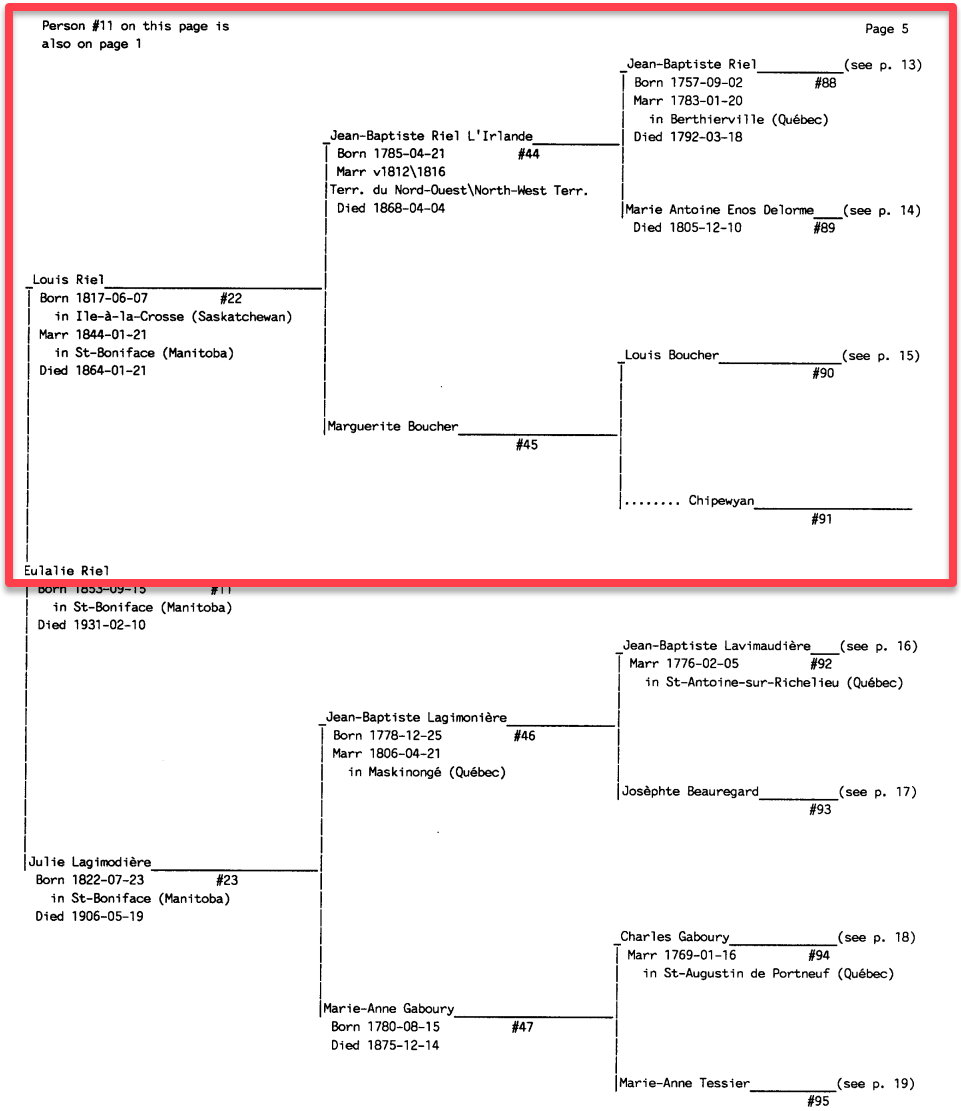
Julie Lagimodière
Born 1822-07-23 #23
in St-Boniface (Manitoba)
Died 1906-05-19

**Great-Great-Grand-Uncle Louis**

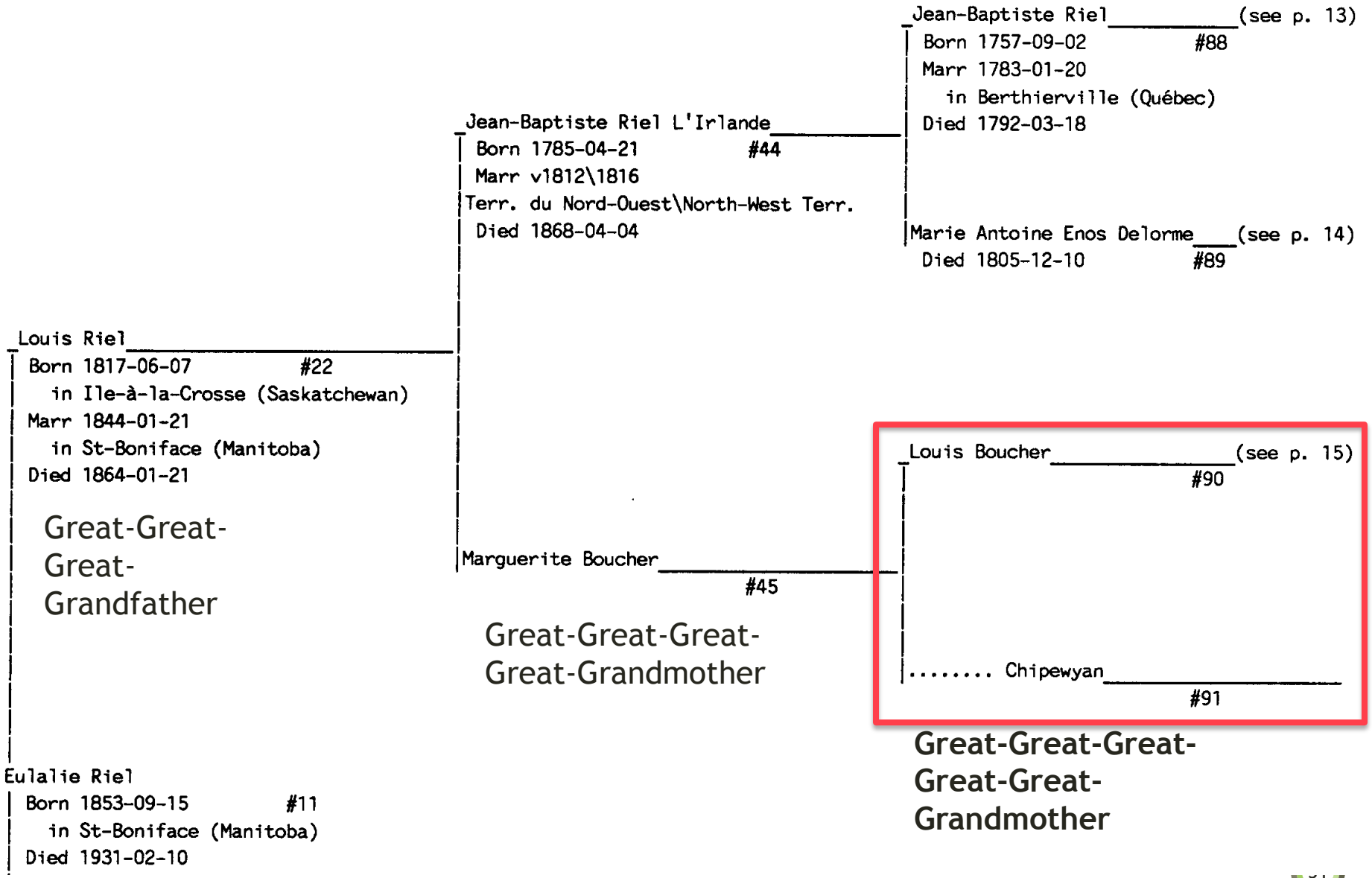
- a Canadian politician, founder of the province of Manitoba, and a political leader of the Métis people.
- Sought to preserve Métis rights and culture
- Now considered a folk hero by many.
- Led two rebellions against the government of Canada.
- While a fugitive, elected three times to the House of Commons of Canada
- Came to believe that he was a divinely chosen leader and prophet
- Arrested, convicted and hanged for high treason

# To the records office…

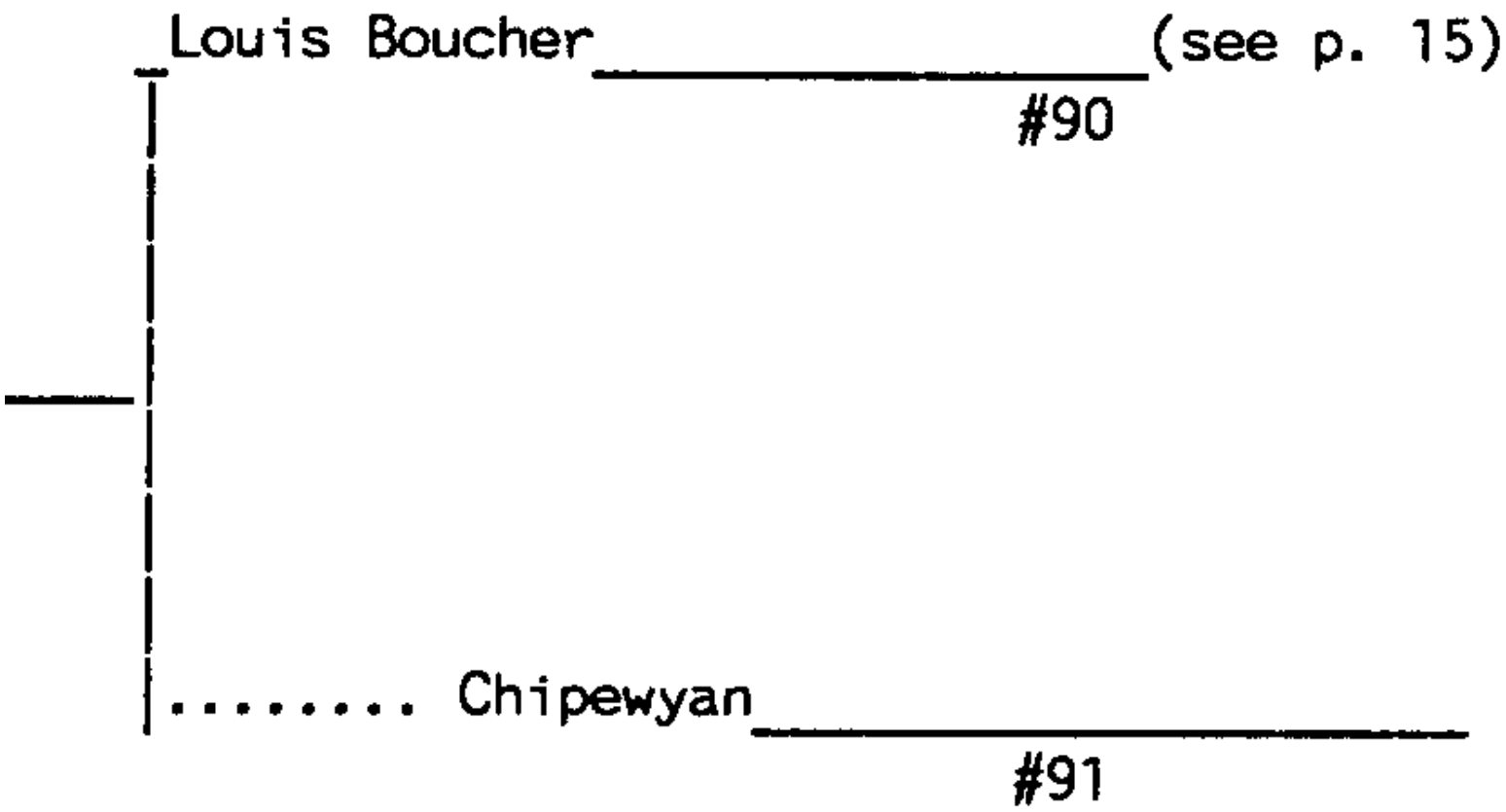
Great-Great-Grandmother



# To the records office...

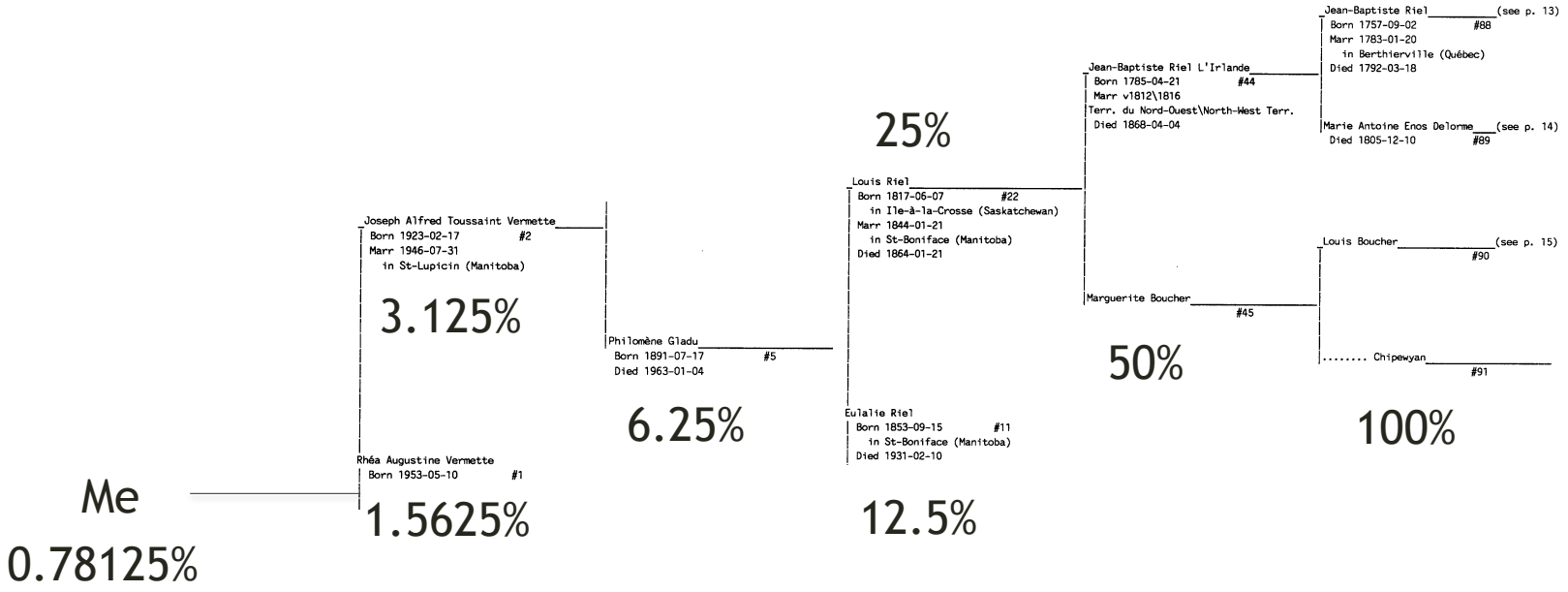


# To the records office…



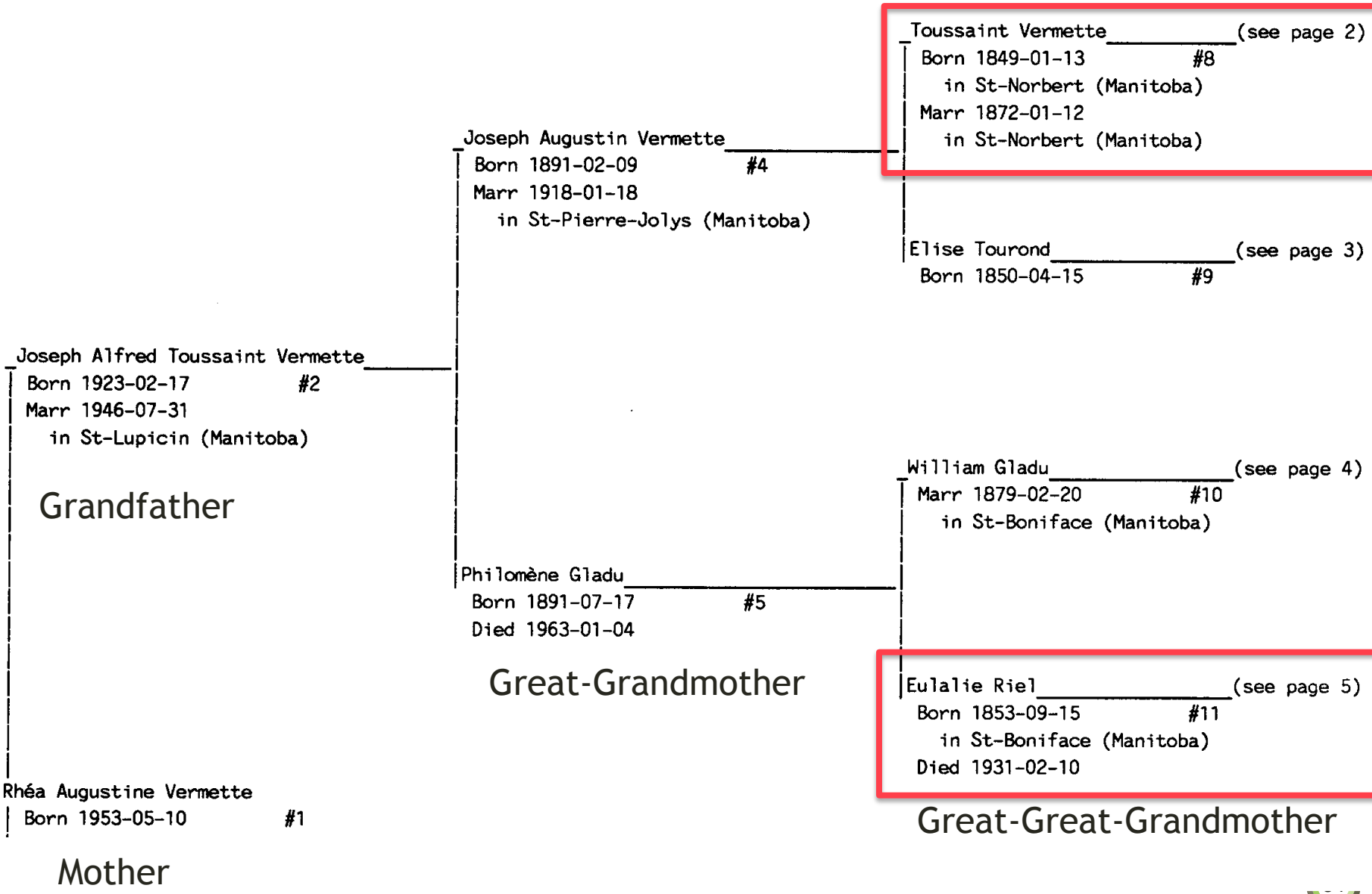

**Great-Great-Great-Great-Great-Great-Grandmother**

# My calculations (~0.8%) substantially less than 23andMe at 3.3%. Why the discrepancy?

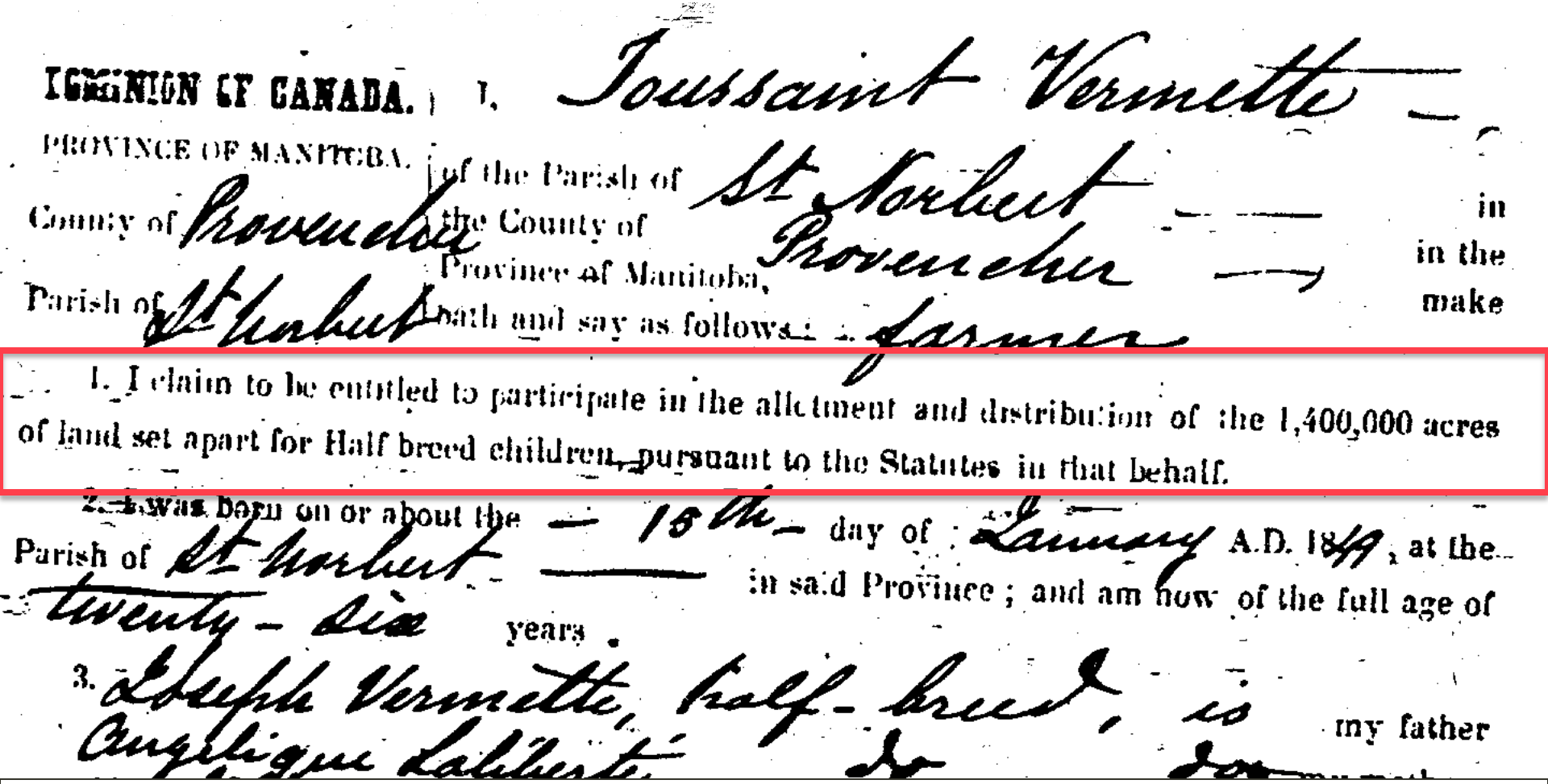



Other sources of Native American Ancestry?

# To the records office...



# Records show at least one other great-great-grandfather had native ancestry



I, Toussaint Vermette

PROVINCE OF MANITOBA, of the Parish of St Norbert in

County of Provencher, the County of Provencher in the

Province of Manitoba, in the

Parish of St Norbert, make oath and say as follows: 1. farmer

1. I claim to be entitled to participate in the allotment and distribution of the 1,400,000 acres of land set apart for Half breed children pursuant to the Statutes in that behalf.

2. I was born on or about the 18th day of January A.D. 1849, at the Parish of St Norbert in said Province; and am now of the full age of twenty-six years.

3. Adolphe Vermette, half-breed, is my father

Angelique Laliberté do do my mother

"I claim to be entitled to participate in the allotment and distribution of the 1,400,000 acres of land set apart for the Half breed children pursuant to the Statutes in that behalf"

# Comparing genes between Obi & Malachi: surprise - identical twins!

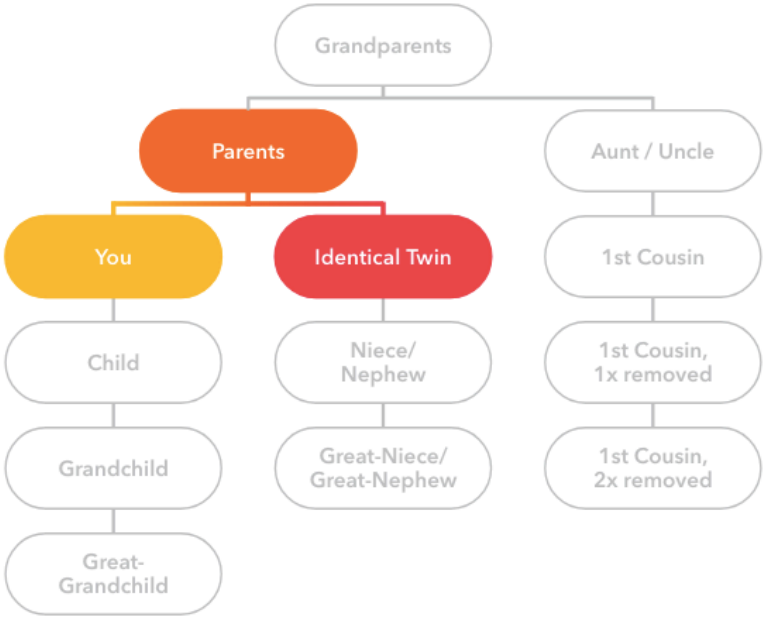
Relationship

We predict Malachi Griffith is your Identical Twin.

Edit relationship

You share **100% of your DNA** with Malachi. View your shared DNA

**You and your twin don't just share the same ancestors – you also share all of your DNA!**

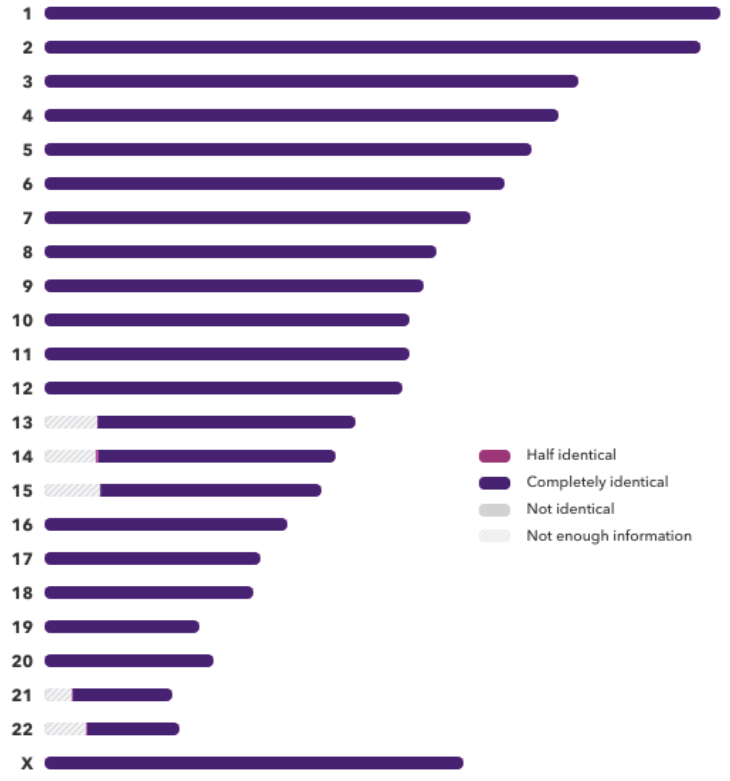



You share **100%** of your DNA with Malachi Griffith

**Your identical DNA is found in 46 segments.**

Since you have two copies of each chromosome, 23andMe can determine whether you share identical segments of DNA on one or both copies of each chromosome.

Learn more about comparing DNA segments.



# Independent test: Monozygosity confirmed!

GFI LABORATORY

681 Main Street • Rockwood, PA 15557 • 814-926-2426

**Twin Zygosity Test Report**

| Date Collected | Case ID# | NAME | Relationship | Sample Type |
|---|---|---|---|---|
| 4/25/2011 | GF12808T1 | Obi Griffith | Twin 1 | Standard |
| 4/25/2011 | GF12808T2 | Malachi Griffith | Twin 2 | Standard |

**DNA Test Data:**

| Genetic System | GF12808T1 | GF12808T2 | Outcome |
|---|---|---|---|
| D8S1179 | 10,15 | 10,15 | Match |
| D21S11 | 29,33.2 | 29,33.2 | Match |
| D7S820 | 9,11 | 9,11 | Match |
| CSF1PO | 11,11 | 11,11 | Match |
| D3S1358 | 16,16 | 16,16 | Match |
| TH01 | 9.3,9.3 | 9.3,9.3 | Match |
| D13S317 | 8,11 | 8,11 | Match |
| D16S539 | 12,13 | 12,13 | Match |
| D2S1338 | 23,23 | 23,23 | Match |
| D19S433 | 12,13 | 12,13 | Match |
| vWA | 15,16 | 15,16 | Match |
| TPOX | 8,9 | 8,9 | Match |
| D18S51 | 17,19 | 17,19 | Match |
| Amelogenin | X,Y | X,Y | Match |
| D5S818 | 11,12 | 11,12 | Match |
| FGA | 22,23 | 22,23 | Match |

**Explanation of Terms:**

Genetic System – Also known as a genetic location or locus, it represents a position on a chromosome that has been found to exhibit a variation of DNA types.

Alleles – The two numbers displayed for a genetic location are known as alleles. These numbers represent the genetic contribution that has been passed from parent to offspring. It is not necessary for two full siblings to share an allele in common at each genetic system.

*Reference:*
*Guidance for Standards for Relationship Testing Laboratories, 9th Edition.*

**Conclusion:**

In order for two siblings to be "identical twins" they must share an exact DNA profile. The results of the DNA test indicate that these two individuals share identical DNA profiles. This outcome confirms that these two individuals are monozygotic twins (identical twins).

How was wrong zygosity determined?

Gender – very accurate for diagnosing dizygotic (DZ) if different genders. No value for telling same gender DZ from MZ

Placental examination – MZ have same placenta ~80% of time. DZ twins can appear to have same placenta if similar implantation site.

~15% of identical twins think they are fraternal

# 23andMe allows you to download your raw data

Obi's raw data

| rsid | chromosome | position | Genotype |
|---|---|---|---|
| rs17163588 | 1 | 26322596 | CT |
| rs212964 | 1 | 48493745 | AG |
| rs533437 | 1 | 151386593 | CT |
| rs1336775 | 1 | 176231201 | AA |
| rs7594106 | 2 | 148089910 | AG |
| rs1004368 | 2 | 168350216 | AA |
| rs1560872 | 2 | 178935533 | AA |
| rs990672 | 2 | 229664562 | CT |
| rs7632996 | 3 | 5767135 | AA |
| rs17033759 | 3 | 10993680 | AG |
| rs1866773 | 3 | 29753247 | TT |
| … | … | … | … |

'diff' with Malachi

| rsid | chromosome | position | Obi | Malachi |
|---|---|---|---|---|
| rs2570006 | 3 | 5126637 | GT | -- |
| rs7632996 | 3 | 5767135 | AA | GG |
| rs3804968 | 3 | 7477700 | CC | -- |
| rs9874577 | 3 | 8728915 | -- | AA |
| rs370526 | 3 | 9045495 | CC | -- |
| rs12490065 | 3 | 10602561 | -- | AG |
| rs17033759 | 3 | 10993680 | AG | AA |
| rs1797912 | 3 | 12445239 | AC | -- |
| rs7650895 | 3 | 12450162 | -- | AA |
| rs9942108 | 3 | 14990337 | -- | AG |
| rs9825720 | 3 | 15390253 | -- | AG |
| … | … | … | … | … |

- 966977 total SNPs
  - 3242 differences (99.66% concordance)
- 959840 called for both twins
  - 64 called differences (99.993% called concordance)
    - Approx. accuracy of platform