

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.



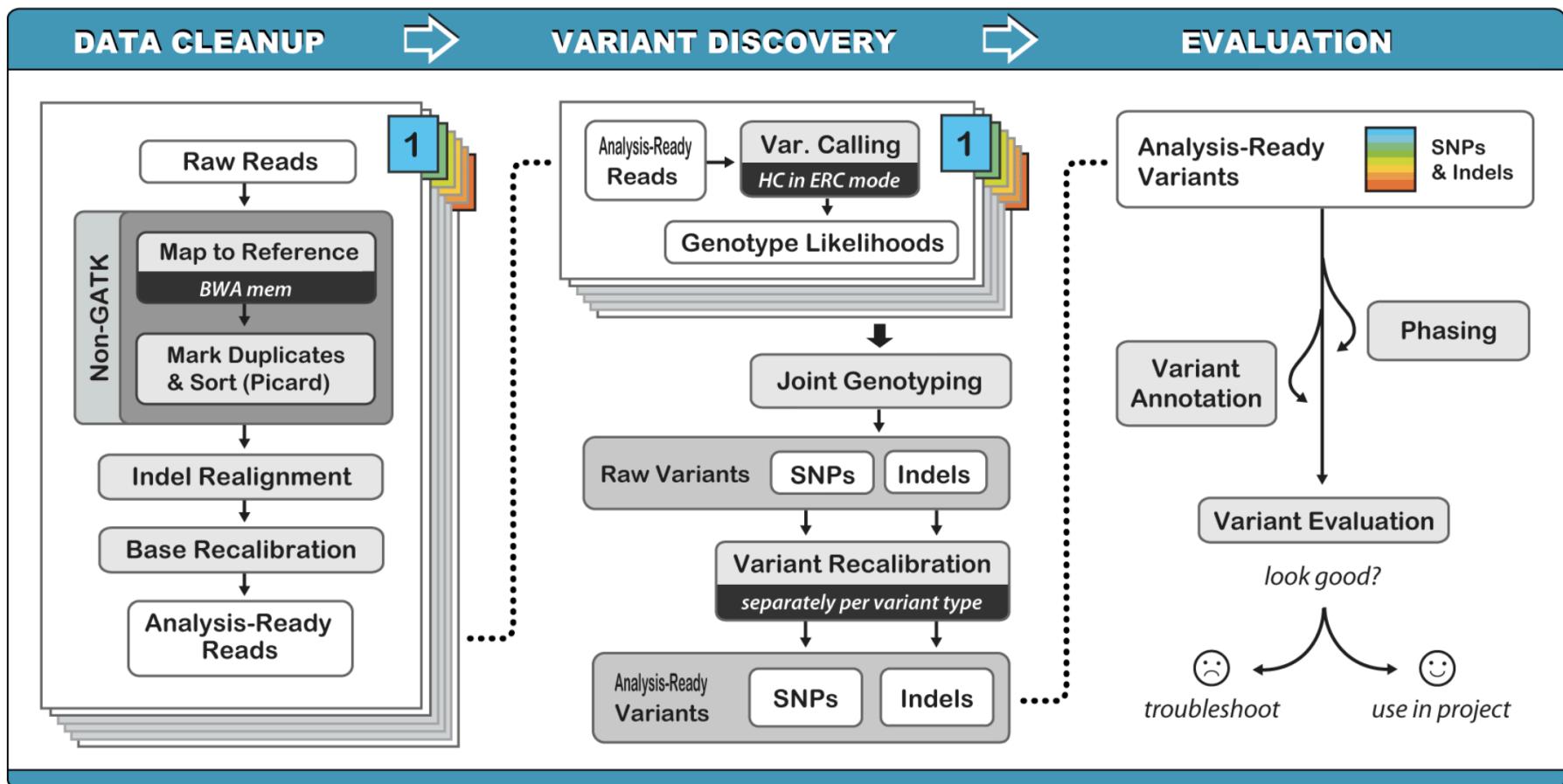
ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

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

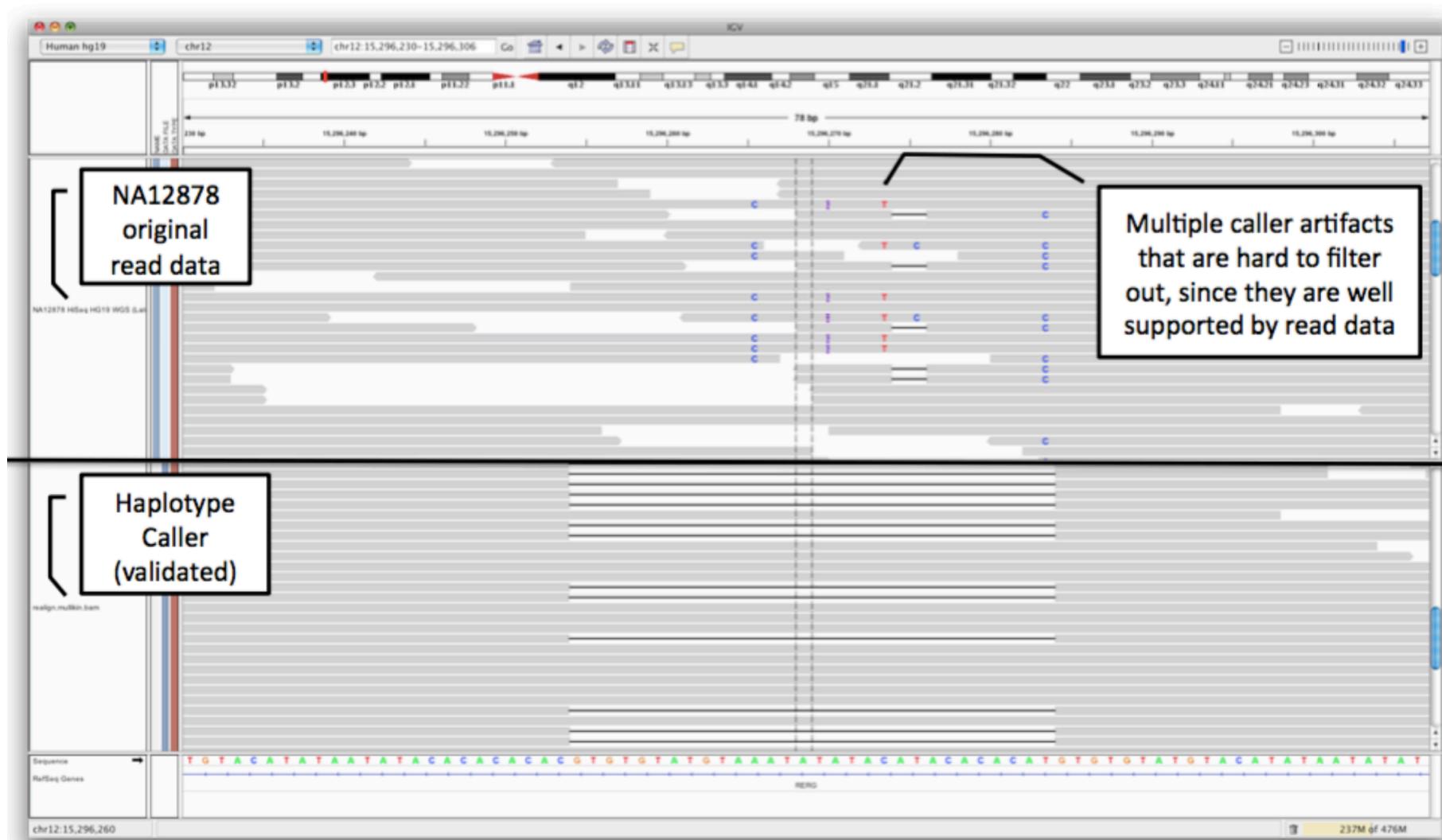
An example germline variant workflow



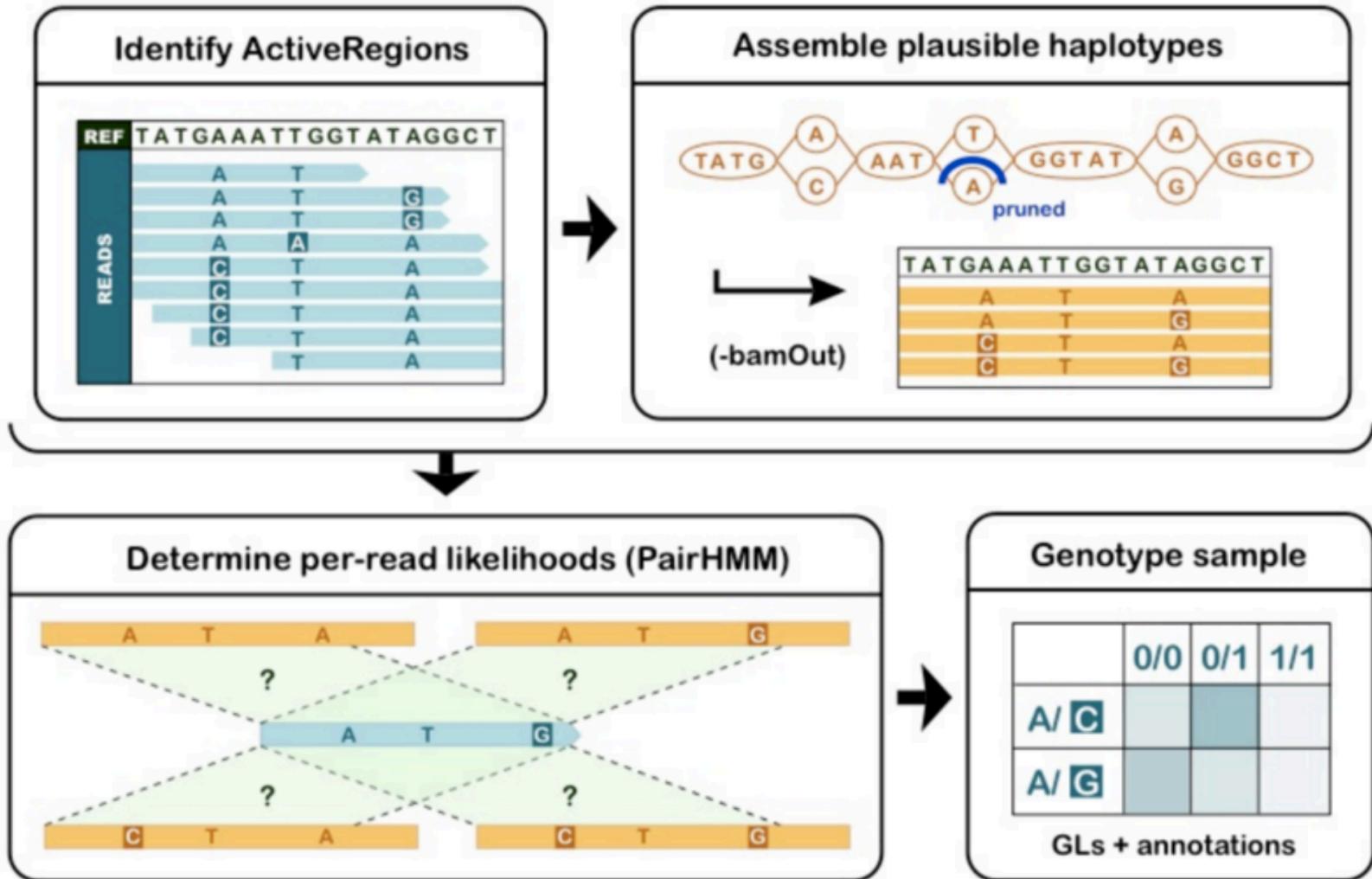
Different methods

- 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

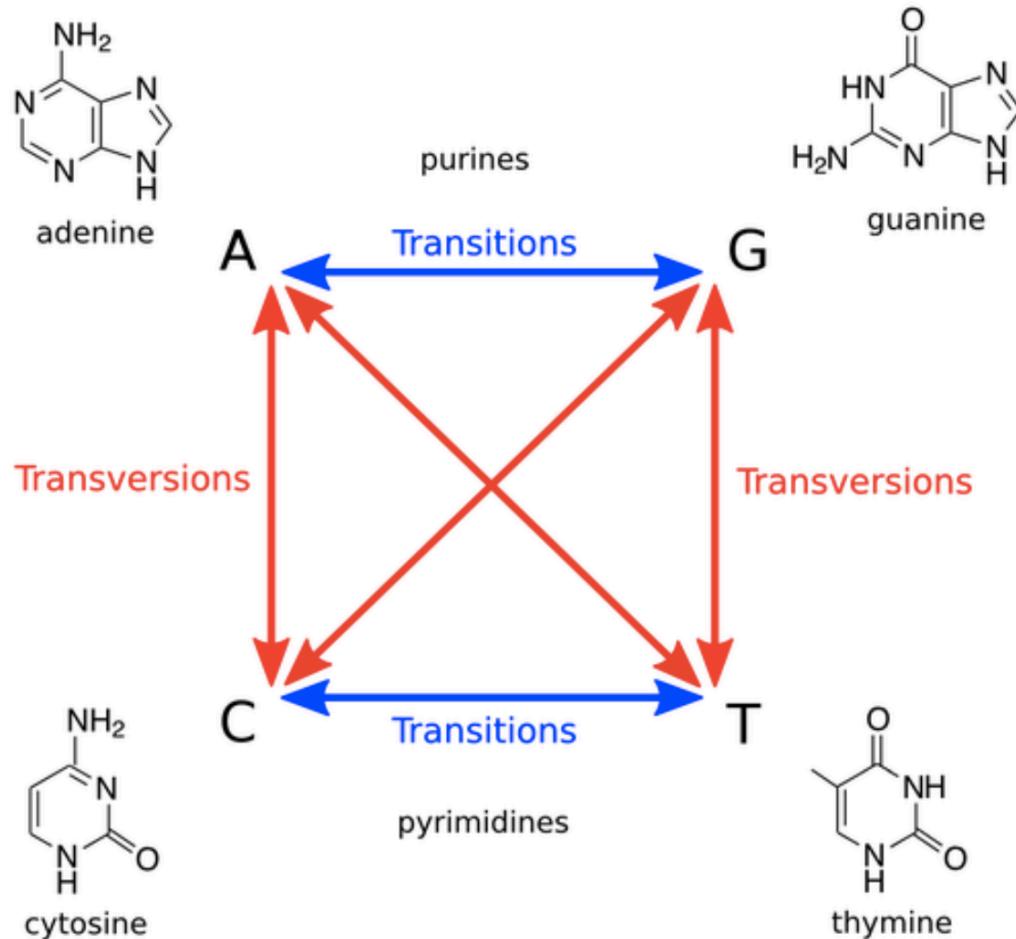
- SNPs
 - QD < 2.0 (variant quality/depth of non-ref samples)
 - MQ < 40.0 (Mapping quality)
 - FS > 60.0 (Phred score Fisher's test pvalue for strand bias)
 - SOR > 3.0 (Strand odds ratio, aims to evaluate whether there is strand bias in the data—updated form of FET)
 - MQRankSum < -12.5 (mapping quality 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)



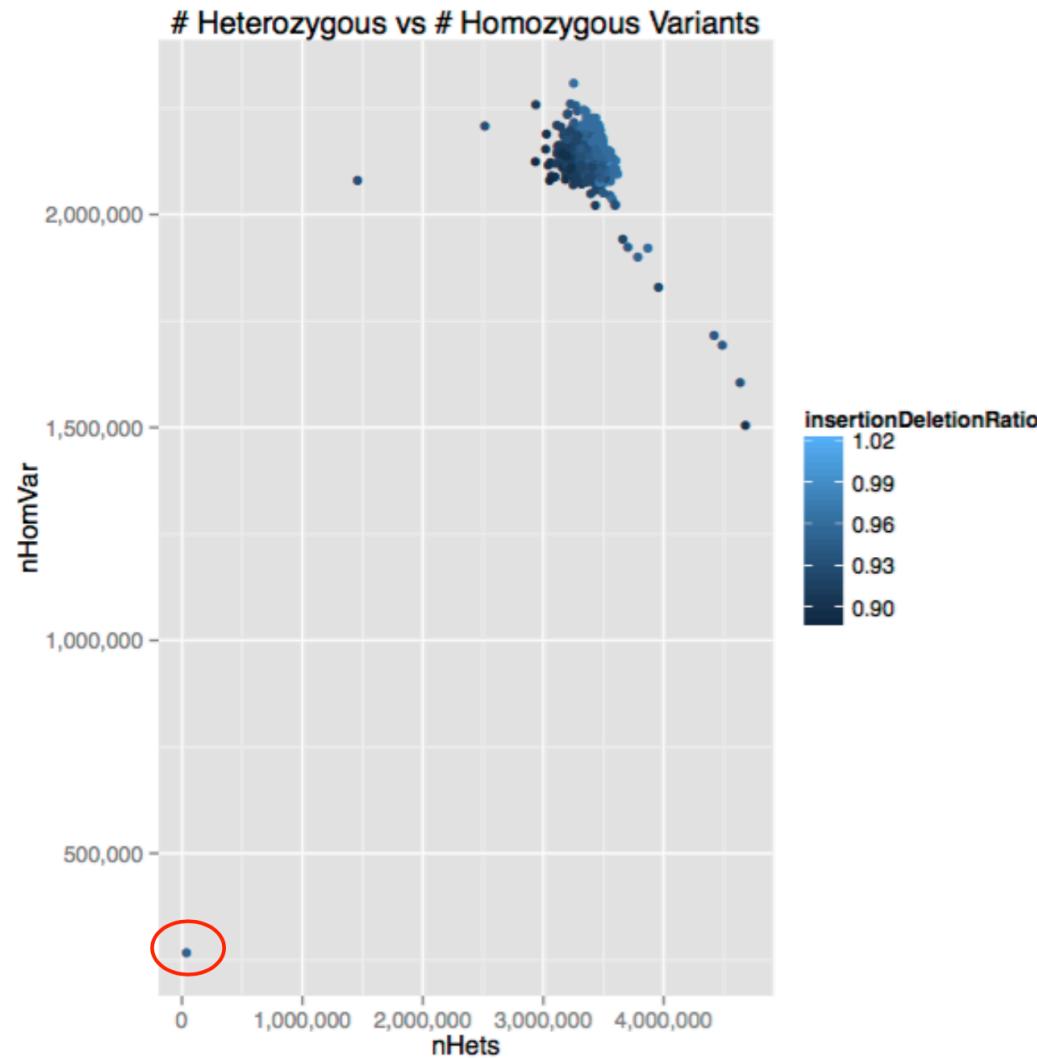
Random = 0.5

WGS = 2.0-2.1

Exome = 3-3.5

Watch for major deviation from typical ratio

heterozygous and homozygous variant



Watch for outliers ...

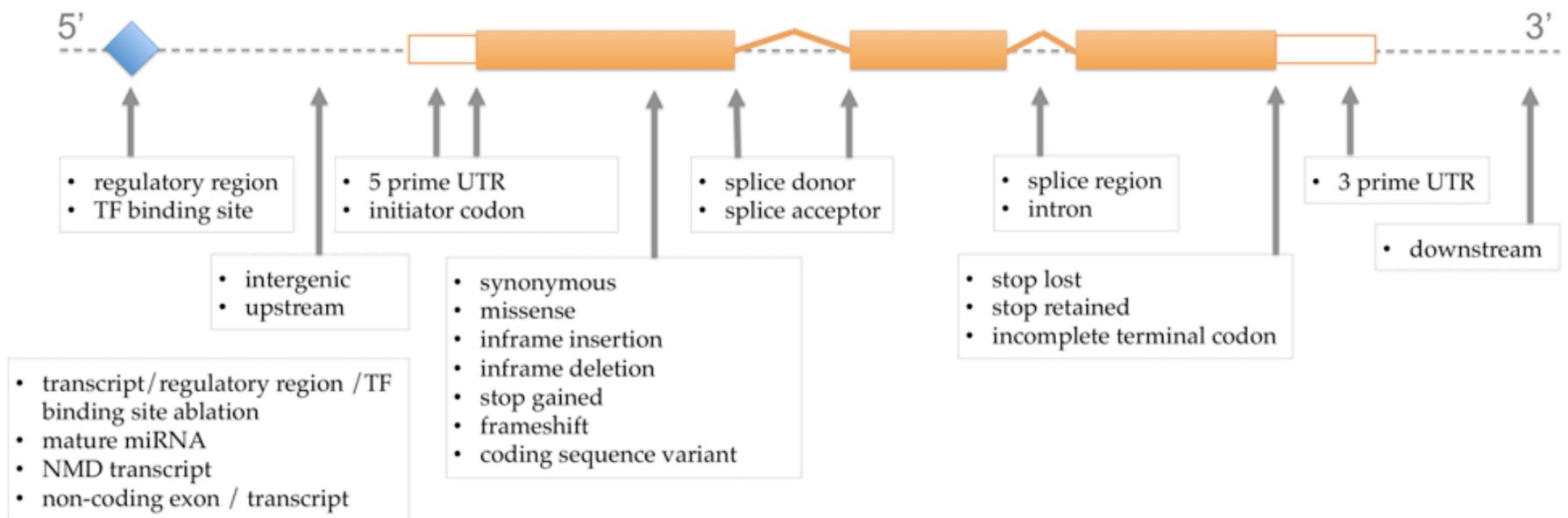
Other things to evaluate

- 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
 - SIFT predicts whether an amino acid substitution affects 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
 - Combined Annotation Dependent Depletion
 - 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?

The screenshot shows the 23andMe website's "SHOP" section. It features two main service offerings:

- Ancestry Service**: Priced at \$99. The description states: "Experience your ancestry in a new way! Get a breakdown of your global ancestry by percentages, connect with DNA relatives and more." A green "add to cart" button is available.
- RECOMMENDED Health + Ancestry Service**: Priced at \$199. The description states: "Get an even more comprehensive understanding of your genetics. Receive 75+ online reports on your ancestry, traits and health - and more." A red "add to cart" button is available.

- 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

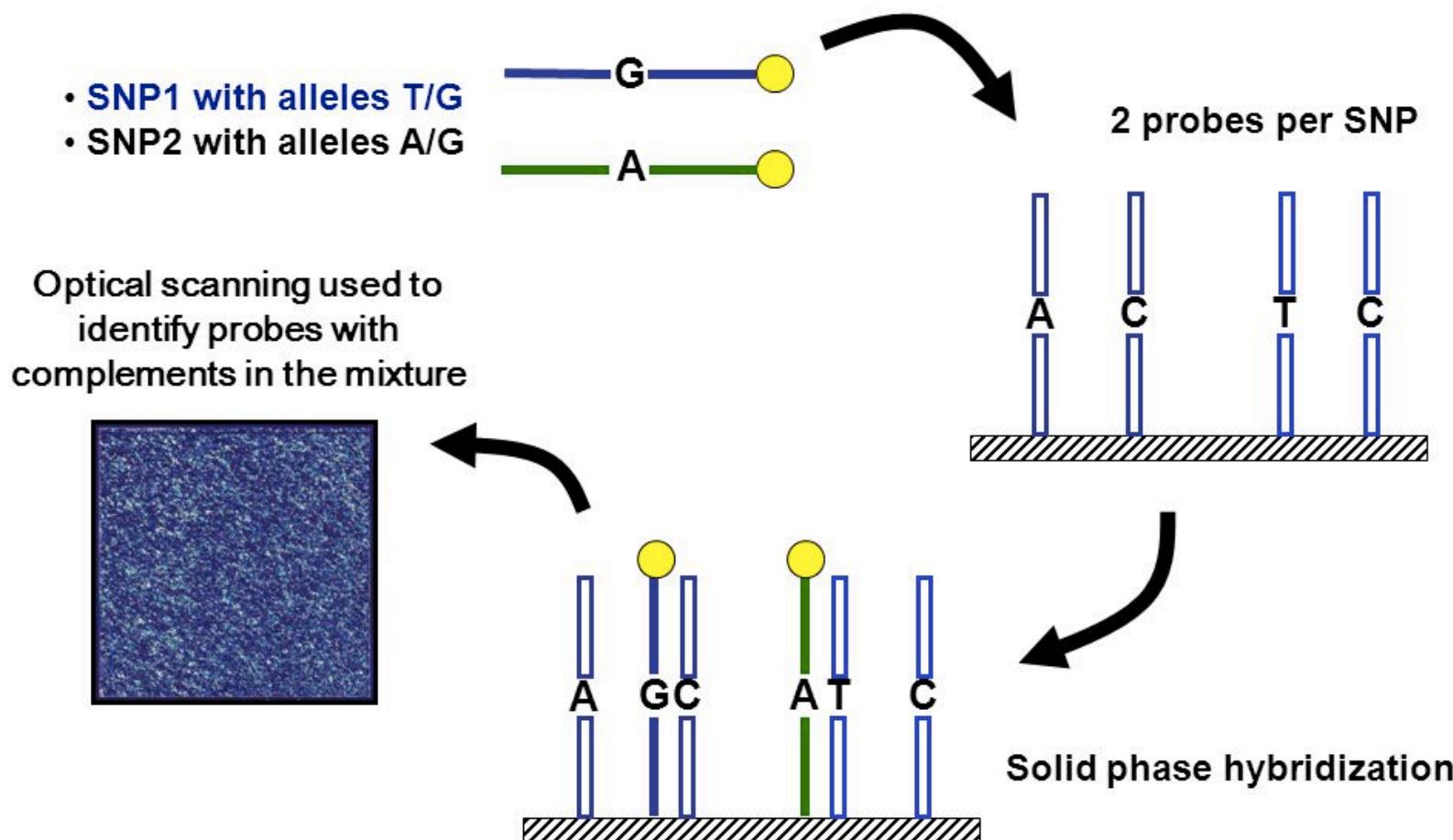
- Primarily genotyping but increasingly also some exome, whole genome, transcriptome and other options
 - Ancestry
 - Genotype-phenotype associations
 - Health risks
 - Early disease detection - intervention
 - Undiagnosed disease
 - Family planning
 - Precision medicine
- Representative companies:
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

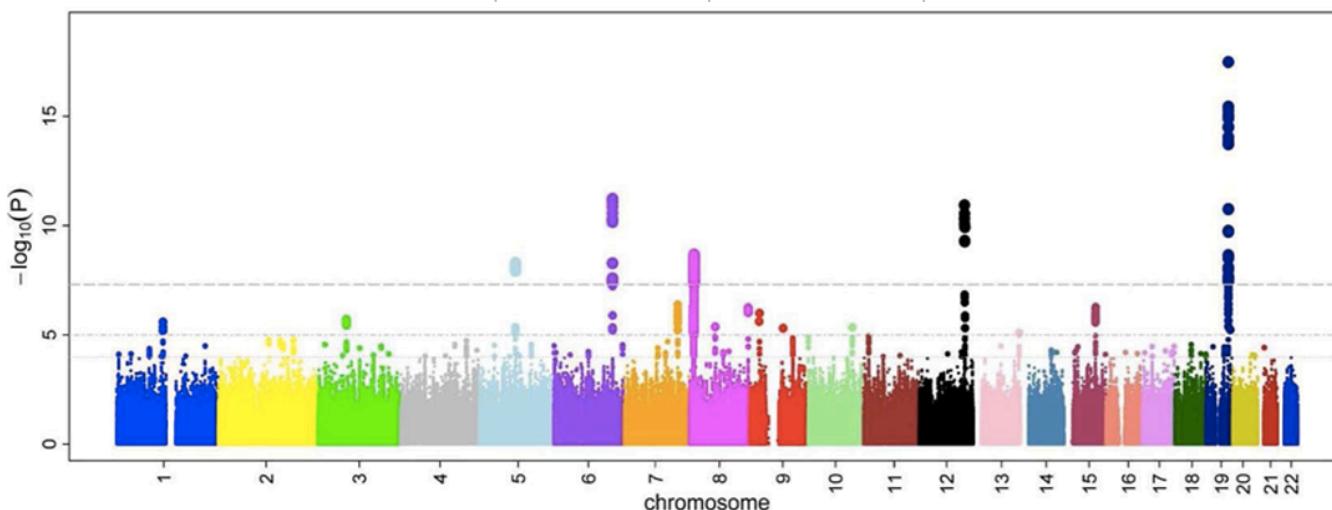
SNP genotyping via direct hybridization



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

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



Search

Obi Griffith | Account ▾ | Help ▾ | Blog | Log out

My Home
Inbox (6)

My Health

Disease Risk
Carrier Status
Drug Response
Traits
Health Labs

My Ancestry
Maternal Line
Paternal Line
Relative Finder
Ancestry Painting
Global Similarity
Ancestry Labs

Sharing & Community
Compare Genes
Family Inheritance
23andMe Community
Genome Sharing

23andWe
Research Surveys (29)
Research Snippets
Research Initiatives
Research Discoveries

Health risks
Traits, etc

Ancestry

Sharing
Comparing



welcome to you.



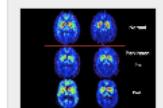
New Blog Post: [Cinco-de-MeO - 23andMe Turns Five!](#)

The typical five year-old is pretty confident, filled with energy and imagination. Maybe a little loud. Yesterday, 23andMe turned five. We're brimming with vigor and vision. Still a little brash. Conceived with the goal of allowing people to explore their genetic information, 23andMe uses new tools that allow people to gain insight into their ancestry, genealogy and health. Who knew that would be so controversial? [continue reading »](#)

2 days ago



A report was updated: [Parkinson's Disease](#)



This report is now applicable to individuals with Asian ancestry. This report also now includes information on i4000415 (the N370S mutation in the GBA gene), rs356220, and rs393152 applicable to individuals with European ancestry.
[continue reading and learn more »](#)

4 days ago



There is a new report for [Glaucoma: Preliminary Research](#)

Glaucoma is the second most common cause of blindness. It is actually a group of closely related eye conditions – all share the features of damage to the optic nerve accompanied by abnormally high pressure inside the eye. The most common type of ... [continue reading and learn more »](#)

11 days ago



There is a new report for [Caffeine Consumption](#)

Find it hard to keep those eyes open in the morning without that cup of coffee? Always

Potential relatives

Inbox (6)

A potential relative would ... Mar 18
A potential relative would ... Mar 17
A potential relative would ... Mar 17

[» view all](#)

Research Surveys

1% complete! Take a survey to get to 6% complete.

Research Snippets

Are you exposed to second-hand smoke at work?

- Yes
 No
 I'm not sure/I don't know

[submit answer](#) skip

Suggest a question topic.

Surveys

Snippets

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

Traits: A “likely sprinter” with wet earwax

23andMe Search Search

Obi Griffith | Account ▾ | Help ▾ | Blog | Log out

My Home [Inbox \(6\)](#)

My Health [Disease Risk](#) [Carrier Status](#) [Drug Response](#)

▶ Traits [Health Labs](#)

My Ancestry [Maternal Line](#) [Paternal Line](#) [Relative Finder](#) [Ancestry Painting](#) [Global Similarity](#) [Ancestry Labs](#)

Sharing & Community [Compare Genes](#) [Family Inheritance](#) [23andMe Community](#) [Genome Sharing](#)

23andWe [Research Surveys \(29\)](#) [Research Snippets](#) [Research Initiatives](#) [Research Discoveries](#)

traits

Share my health results with family and friends

Show results for Obi Griffith [See new and recently updated reports »](#)

23andWe Discoveries were made possible by 23andMe members who took surveys.

Name	Confidence ▲	Outcome
Alcohol Flush Reaction	★★★★	Does Not Flush
Bitter Taste Perception	★★★★	Unlikely to Taste
Earwax Type	★★★★	Wet
Eye Color	★★★★	Likely Brown
Hair Curl	★★★★	Slightly Curlier Hair on Average
Lactose Intolerance	★★★★	Likely Tolerant
Malaria Resistance (Duffy Antigen)	★★★★	Not Resistant
Male Pattern Baldness	★★★★	Decreased Odds
Muscle Performance	★★★★	Likely Sprinter
Non-ABO Blood Groups	★★★★	See Report
Norovirus Resistance	★★★★	Resistant
Resistance to HIV/AIDS	★★★★	Not Resistant
Smoking Behavior	★★★★	If a Smoker, Likely to Smoke More
Adiponectin Levels	★★★	See Report
Asparagus Metabolite Detection	★★★	Typical Odds of Detecting
Birth Weight	★★★	See Report

“Health Labs”: Extra five pounds not my fault

 Search

Obi Griffith | Account ▾ | Help ▾ | Blog | Log out

My Home

Inbox (6)

My Health

Disease Risk

Carrier Status

Drug Response

Traits

▶ Health Labs

My Ancestry

Maternal Line

Paternal Line

Relative Finder

Ancestry Painting

Global Similarity

Ancestry Labs

Sharing & Community

Compare Genes

Family Inheritance

23andMe Community

Genome Sharing

23andWe

Research Surveys (29)

Research Snippets

Research Initiatives

Research Discoveries

Twenty Three and Me health labs

Genetic Weight Calculator

By: Arnab Chowdry, John Garcia

See how much of your weight you can blame on your genes (not your jeans)!

- [Discuss this feature](#)
- [Send feedback](#)

Your genes influence your health and your appearance, so why not your weight as well? Two recent studies genotyped nearly 100,000 adults of European ancestry and confirmed six independent SNPs strongly associated with differences in body mass index (BMI). Try it out and see how much of your weight you can blame on your genes!

Show data for:

Select your height: "

Enter your weight in pounds (optional):

Obi Griffith's entered data:

Height: 5' 11" Weight: 170.0

Calculated BMI: 23.7

This BMI is considered normal weight

Your SNPs account for at least:

5.09 pounds

(0.71 BMI units)

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

 Search

Obi Griffith | Account ▾ | Help ▾ | Blog | Log out

My Home

Inbox (6)

My Health

Disease Risk

Carrier Status

Drug Response

Traits

Health Labs

My Ancestry

Maternal Line

Paternal Line

Relative Finder

Ancestry Painting

Global Similarity

Ancestry Labs

Sharing & Community

Compare Genes

Family Inheritance

23andMe Community

Genome Sharing

23andWe

disease risk

Share my health results with family and friends

Show results for

Obi Griffith

See new and recently updated reports »

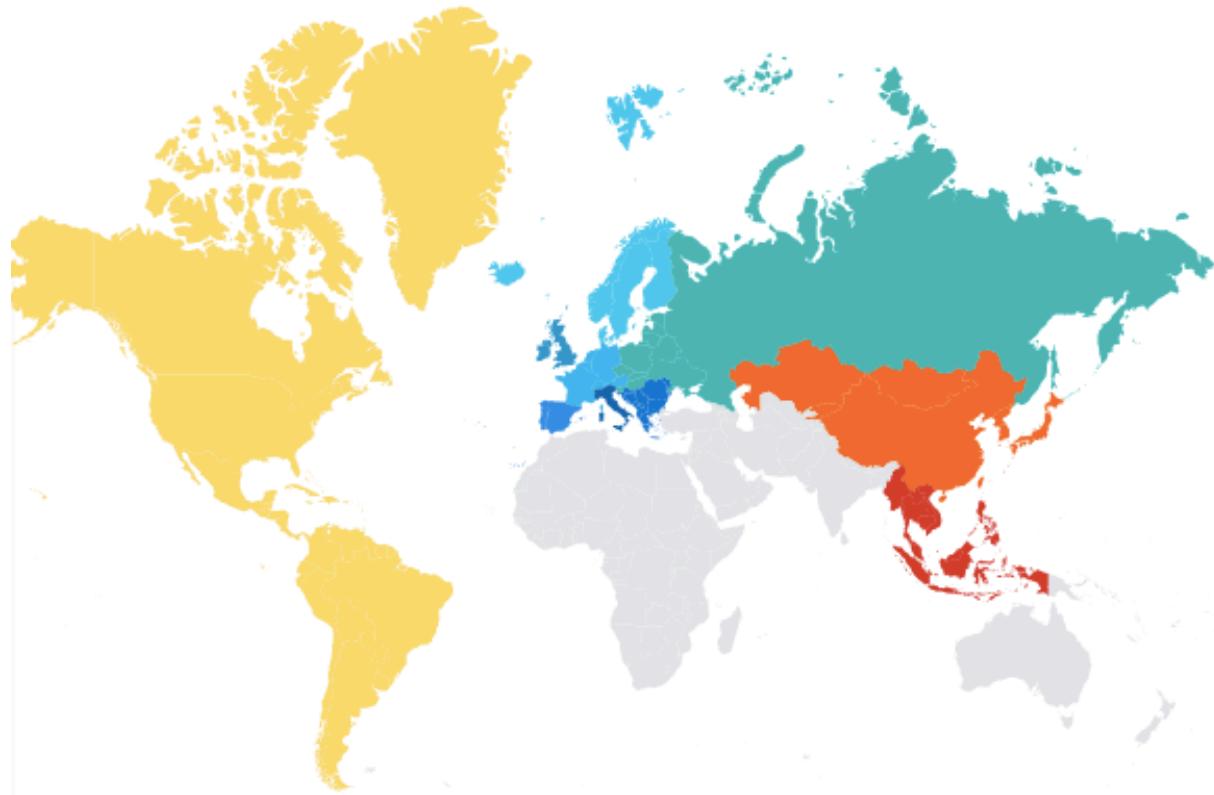
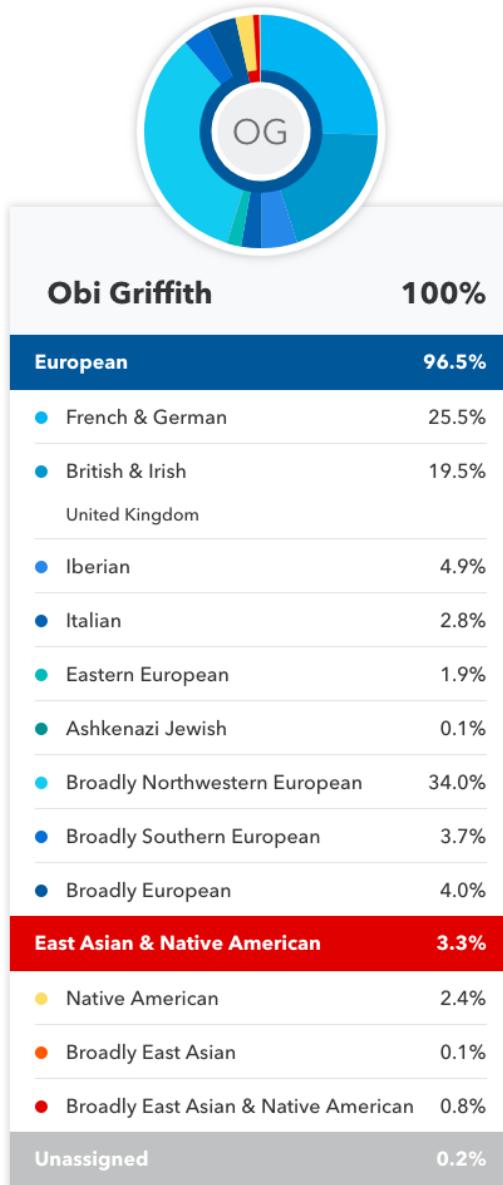
23andWe Discoveries were made possible by 23andMe members who took surveys.

Elevated Risk

Name	Confidence	Your Risk	Avg. Risk	Compared to Average
Type 2 Diabetes	★★★★	32.9%	25.7%	1.28x
Psoriasis	★★★★	16.8%	11.4%	1.48x
Alzheimer's Disease <small>new</small>	★★★★	14.2%	7.2%	1.98x
Age-related Macular Degeneration	★★★★	11.3%	7.0%	1.61x
Colorectal Cancer	★★★★	7.0%	5.6%	1.26x
Rheumatoid Arthritis	★★★★	3.9%	2.4%	1.65x
Type 1 Diabetes	★★★★	1.5%	1.0%	1.50x
Esophageal Squamous Cell Carcinoma (ESCC)	★★★★	0.4%	0.4%	1.21x
Stomach Cancer (Gastric Cardia Adenocarcinoma)	★★★★	0.3%	0.2%	1.22x

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)



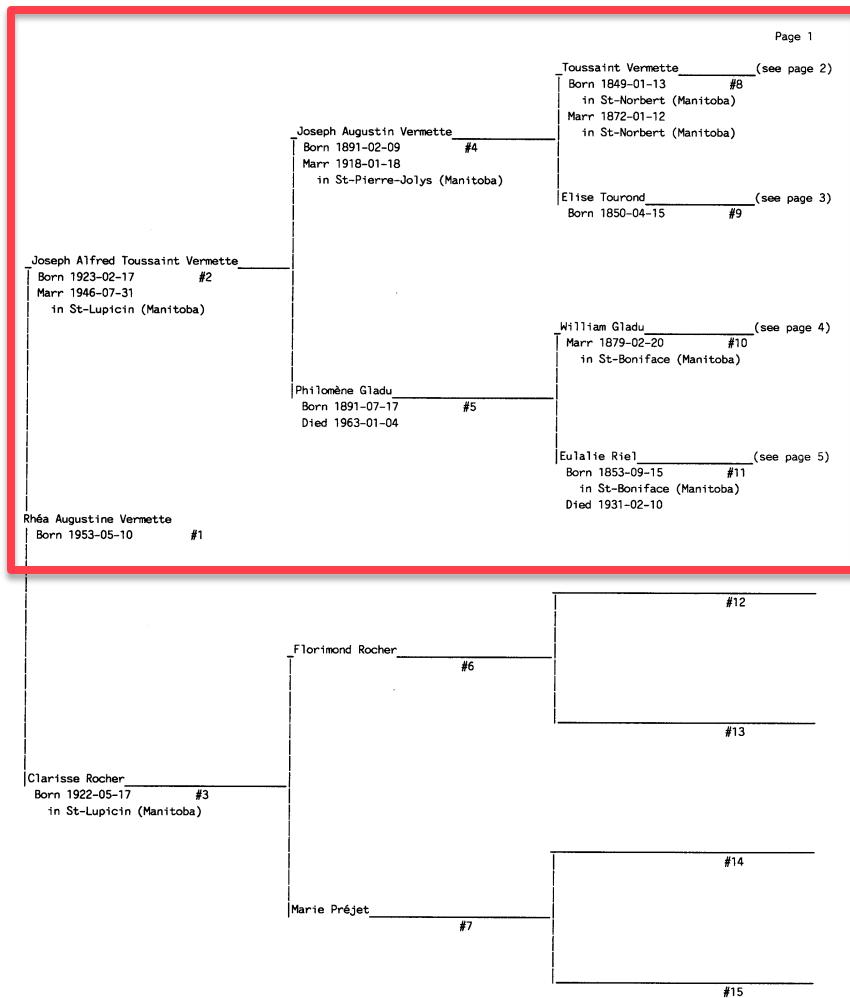
Where did this Asian or Native American component come from?

To the records office...

Grandfather

Mother

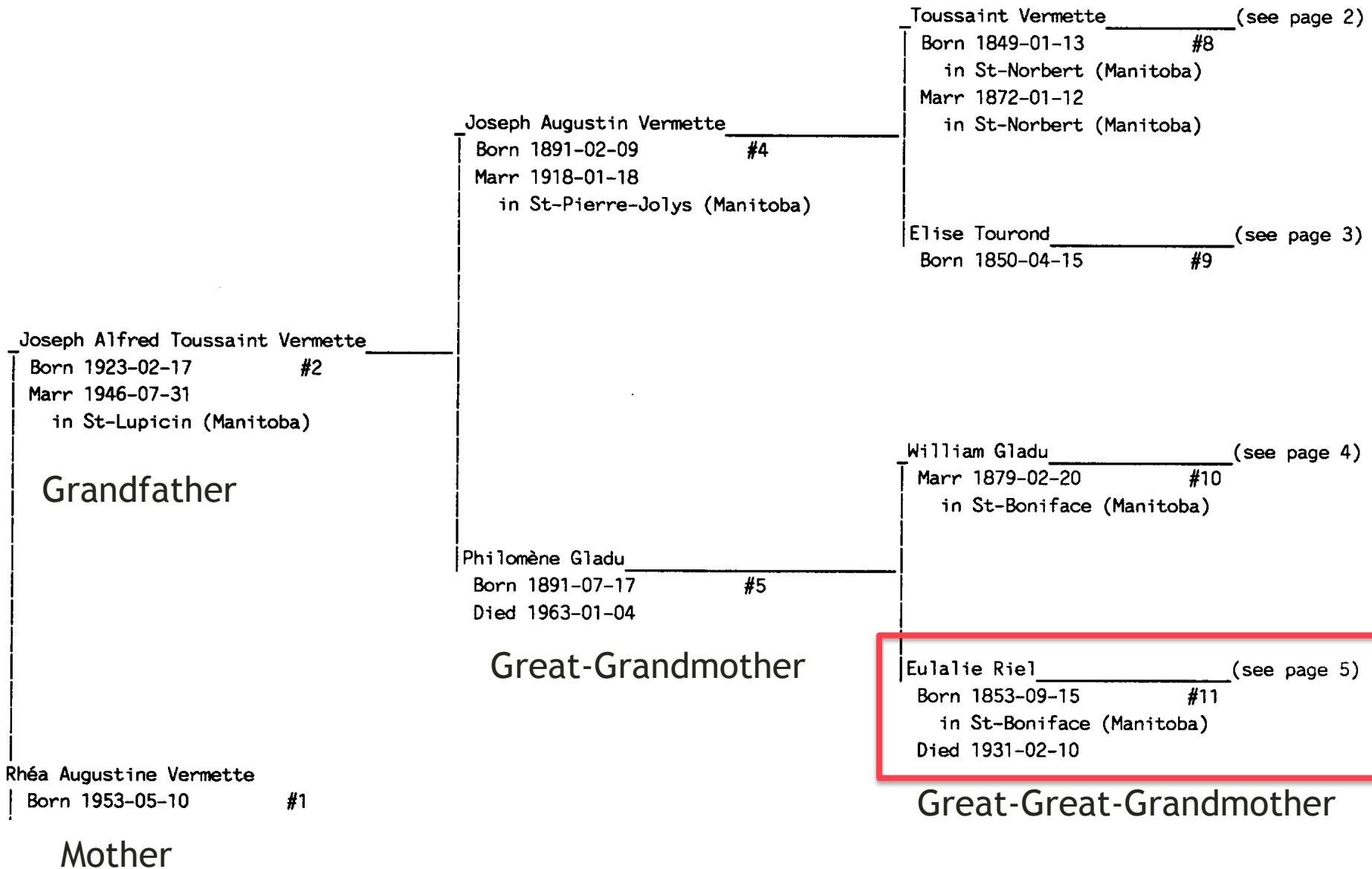
Grandmother



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.

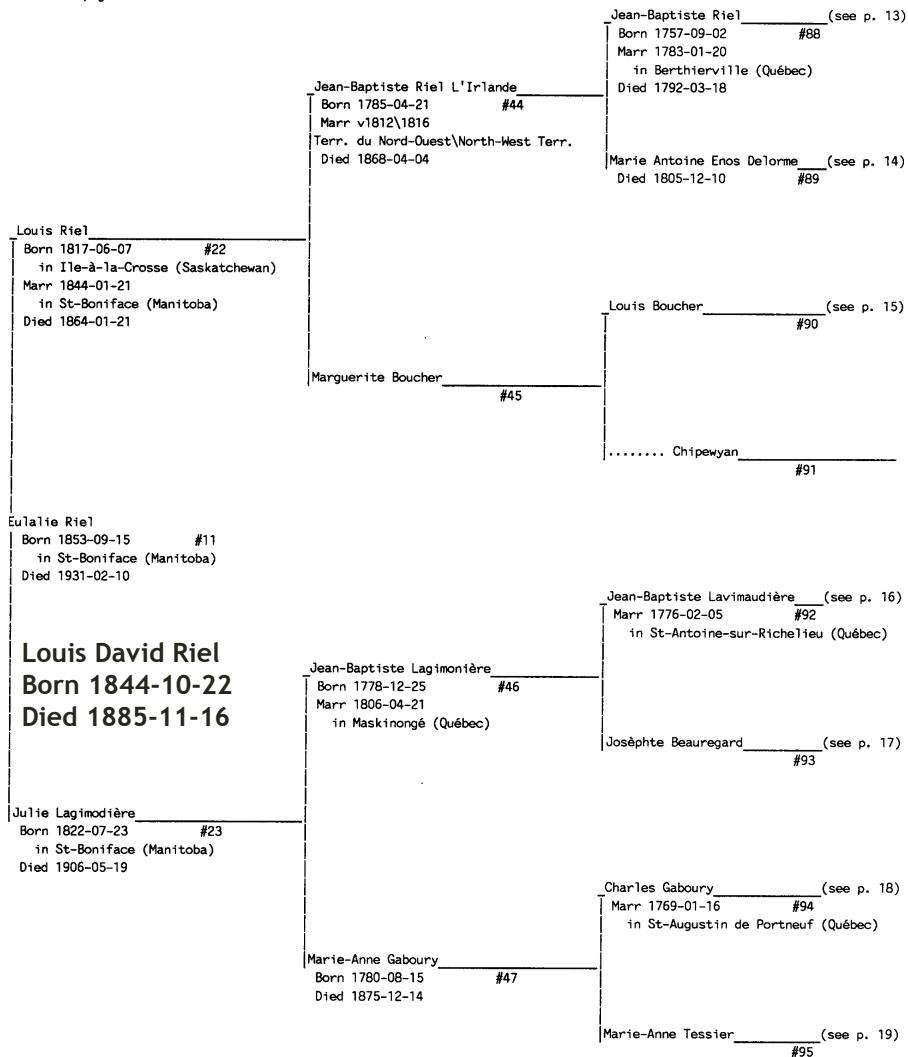
To the records office...



An historical aside...

Person #11 on this page is
also on page 1

Page 5



Great-Great-
Grandmother

An historical aside...

Person #11 on this page is
also on page 1

Great-Great- Grandmother

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

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

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

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

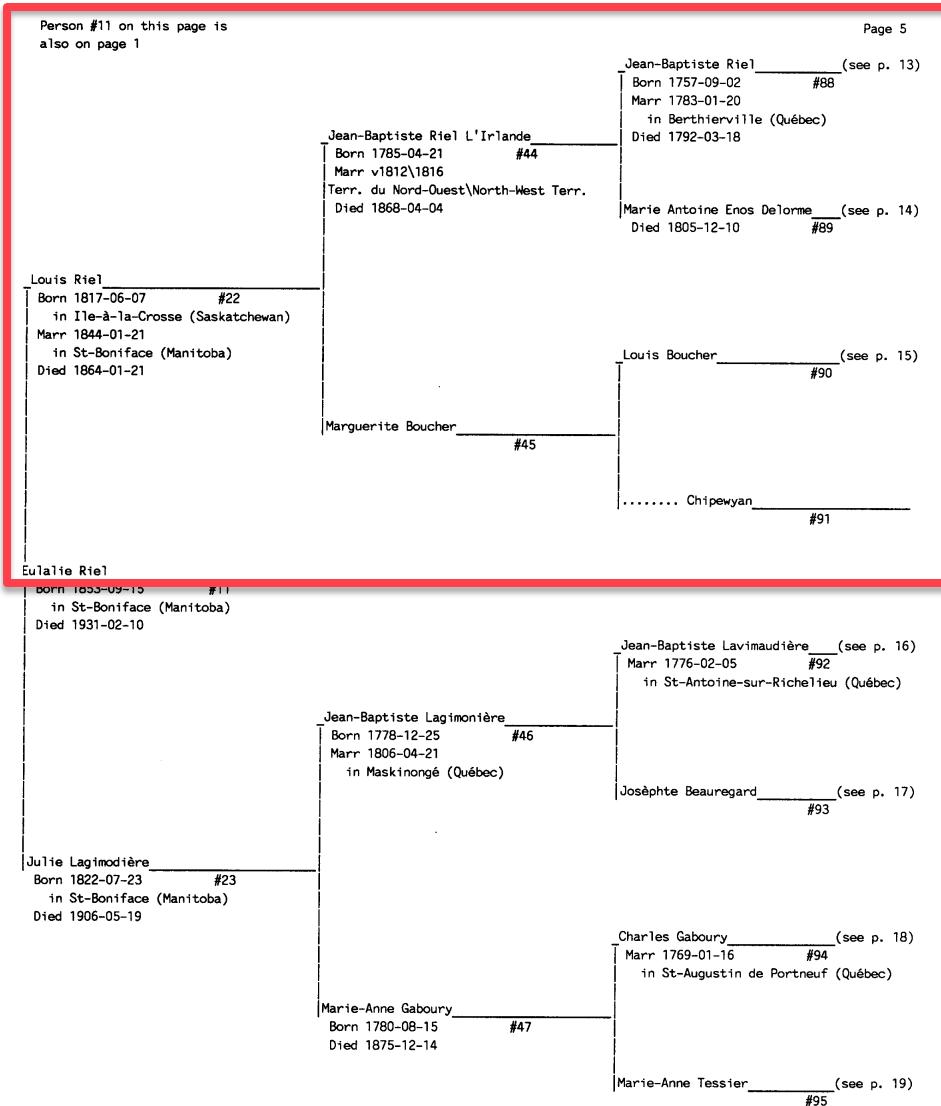


Great-Great-Uncle Louis

- 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...

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

Great-Great-
Great-
Grandfather

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

Great-Great-Grandmother

Jean-Baptiste Riel L'Irlande _____
Born 1785-04-21 #44
Marr v1812\1816
Terr. du Nord-Ouest\North-West Terr.
Died 1868-04-04

Marguerite Boucher _____
#45

Great-Great-Great-
Great-Grandmother

Jean-Baptiste Riel _____ (see p. 13)
Born 1757-09-02 #88
Marr 1783-01-20
in Berthierville (Québec)
Died 1792-03-18

Marie Antoine Enos Delorme _____ (see p. 14)
Died 1805-12-10 #89

Louis Boucher _____ (see p. 15)
#90

..... Chipewyan _____
#91

Great-Great-Great-
Great-Grandmother



To the records office...

Louis Boucher _____ (see p. 15)

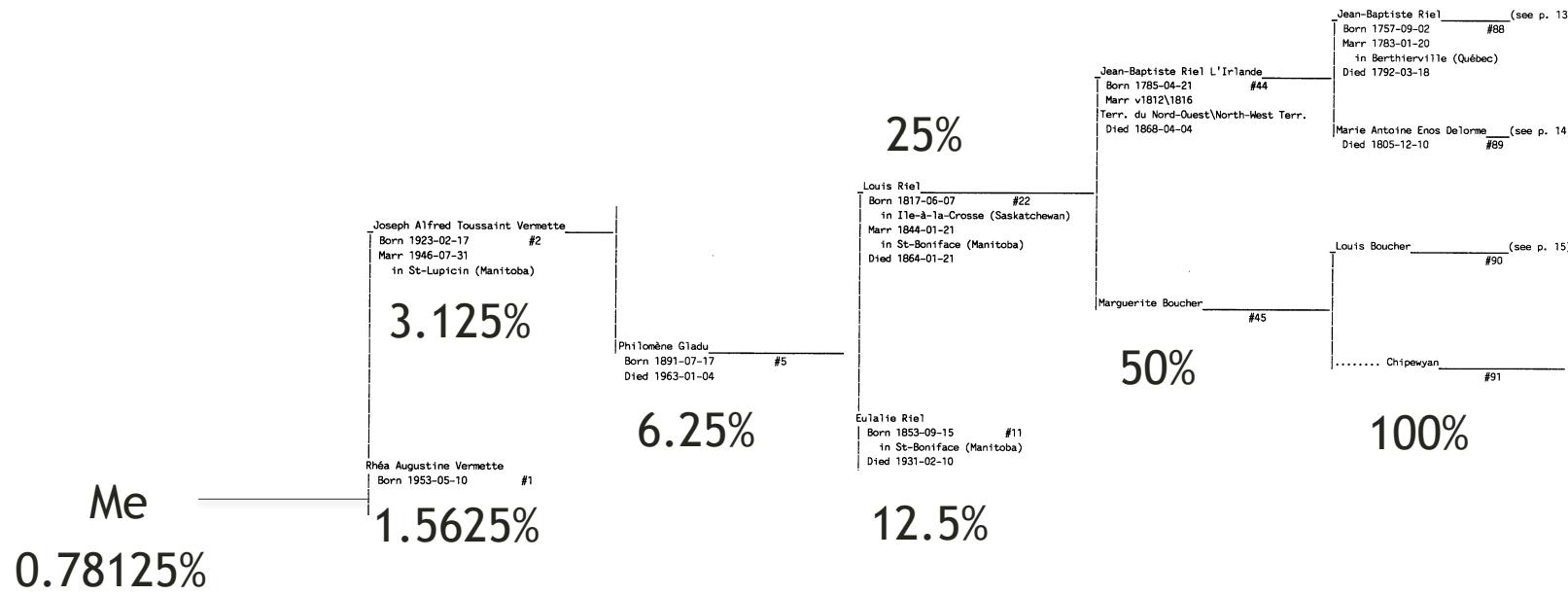
#90

..... Chipewyan _____

#91

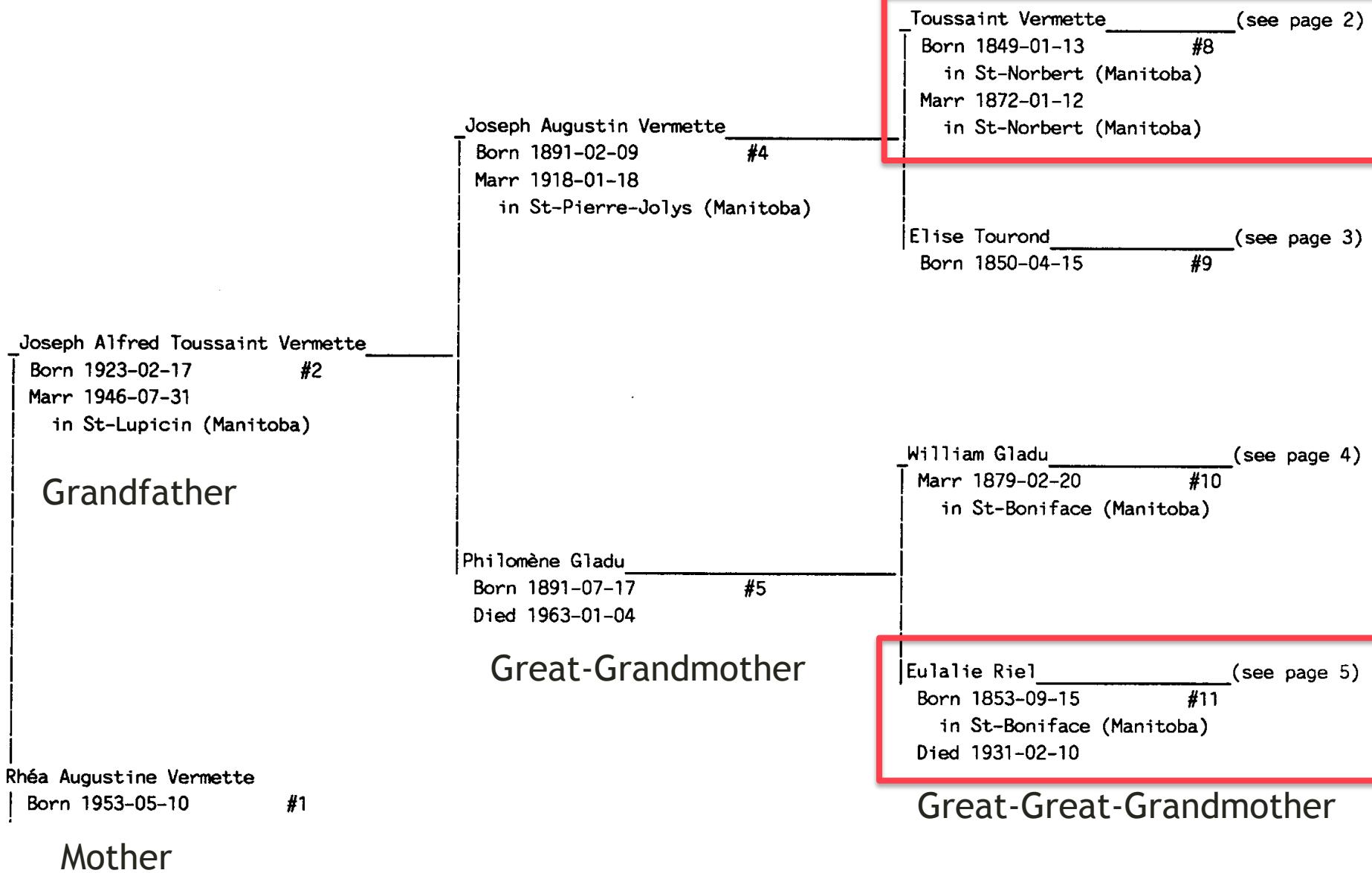
Great-Great-Great-
Great-Great-
Grandmother

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

IMMIGRATION OF CANADA. 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, make
Parish of St. Norbert bath and say as follows: — 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. Joseph Vermette, half-breed, is my father
Anglique Lablanc. do do

“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!

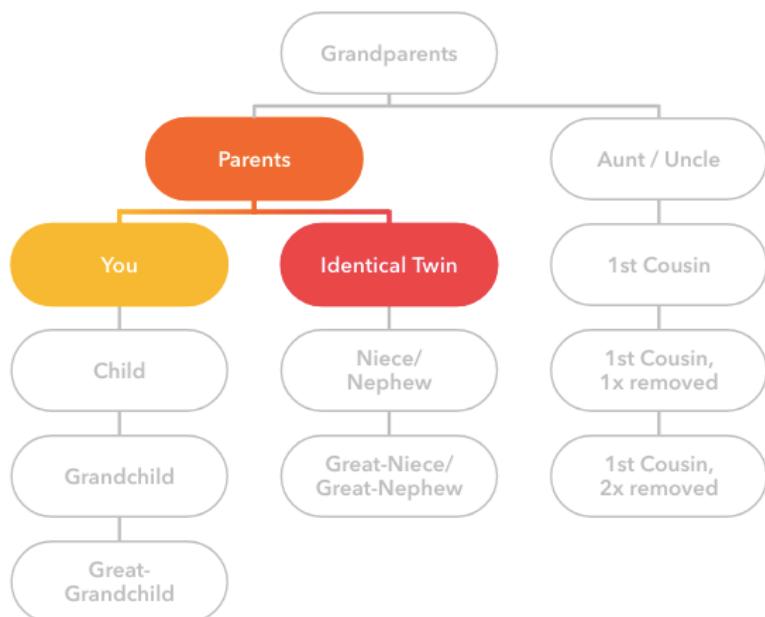
Relationship

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!

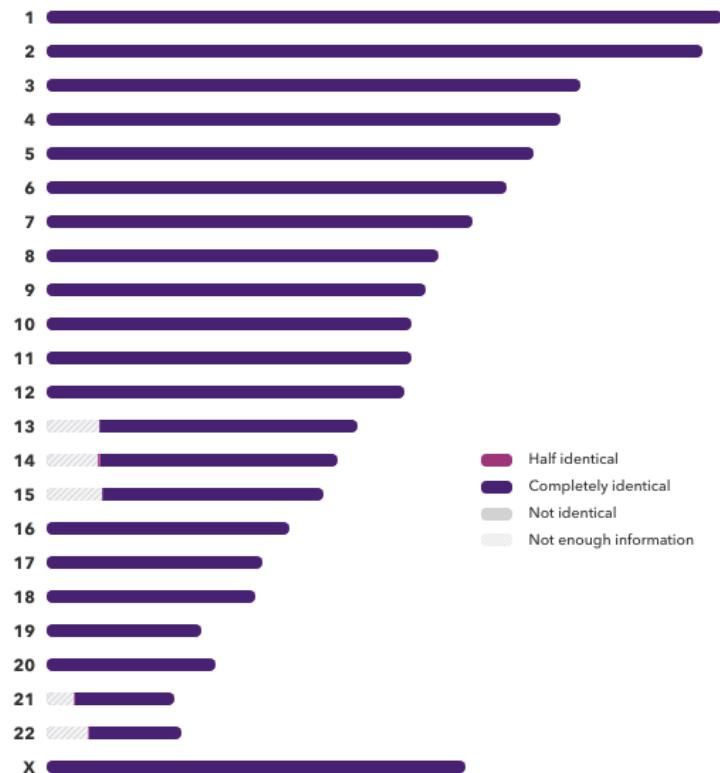


You share **100%** of your DNA with Malachi Griffith

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!



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

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