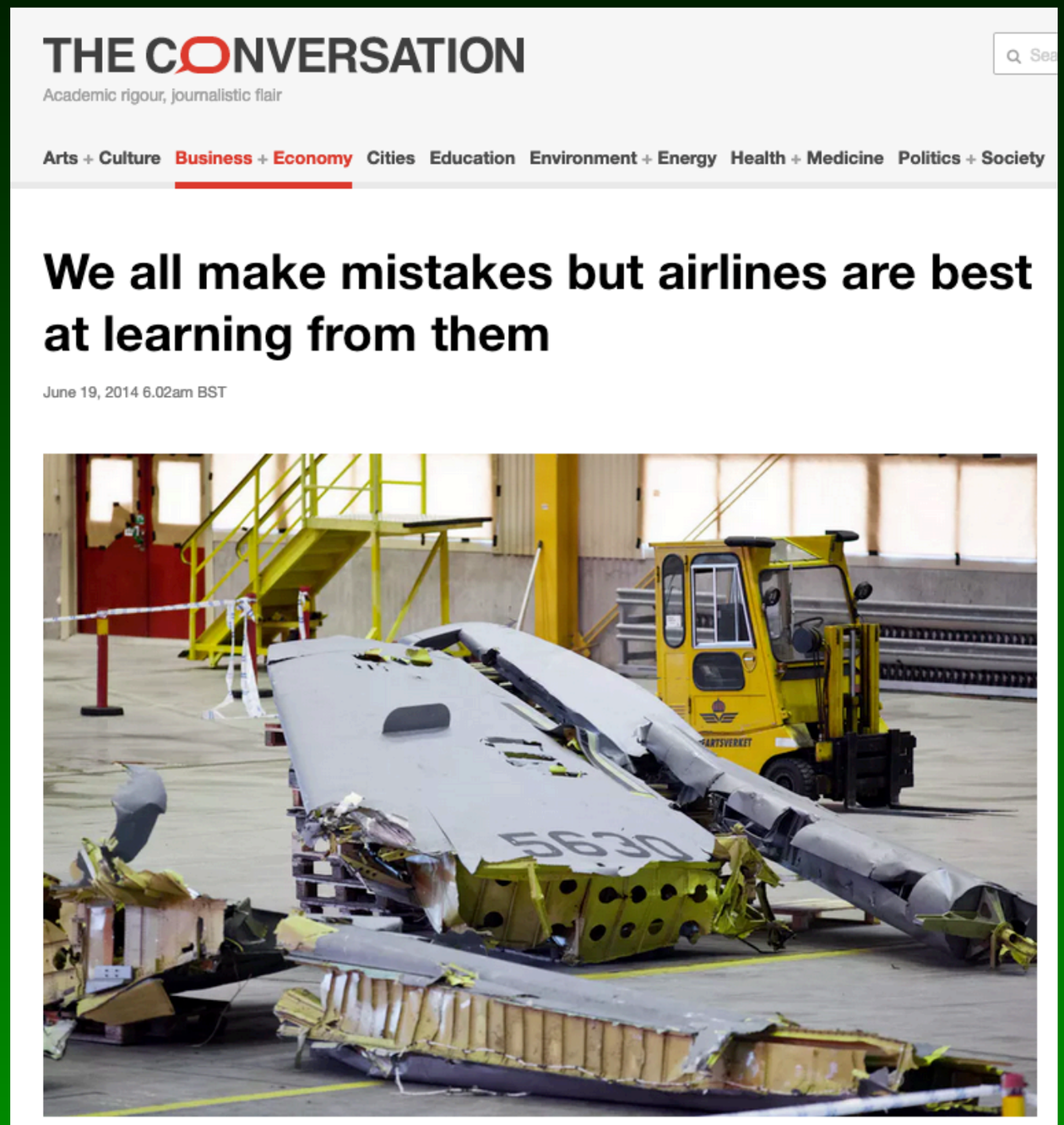
NGS in practice

where things go wrong
(practical experiences)

i.c.w. Anna Benet-Pages
Gijs Santen
Christian Gilissen

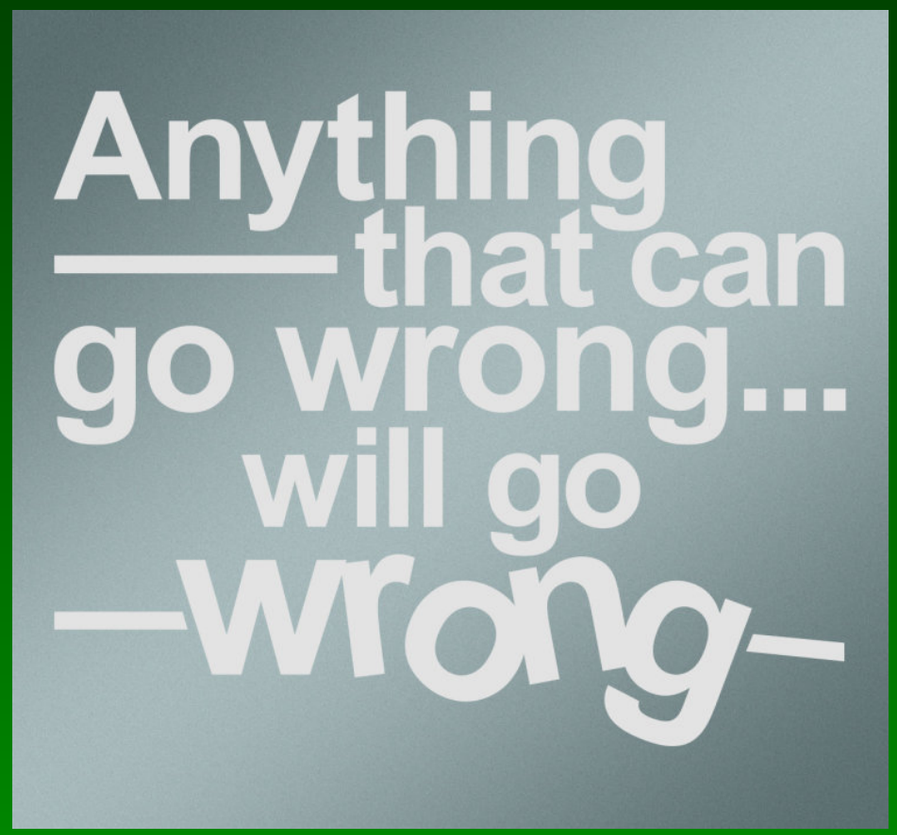
Johan den Dunnen

Accept mistakes



Accept mistakes

- report incidents and near incidents (!)
- perform extensive analysis
- do not focus on blame
- recommend changes



unfortunately

NGS workflow

- sampling
- sample preparation
- sequencing
- mapping reads
- call variants
- annotate variants
- variant interpretation

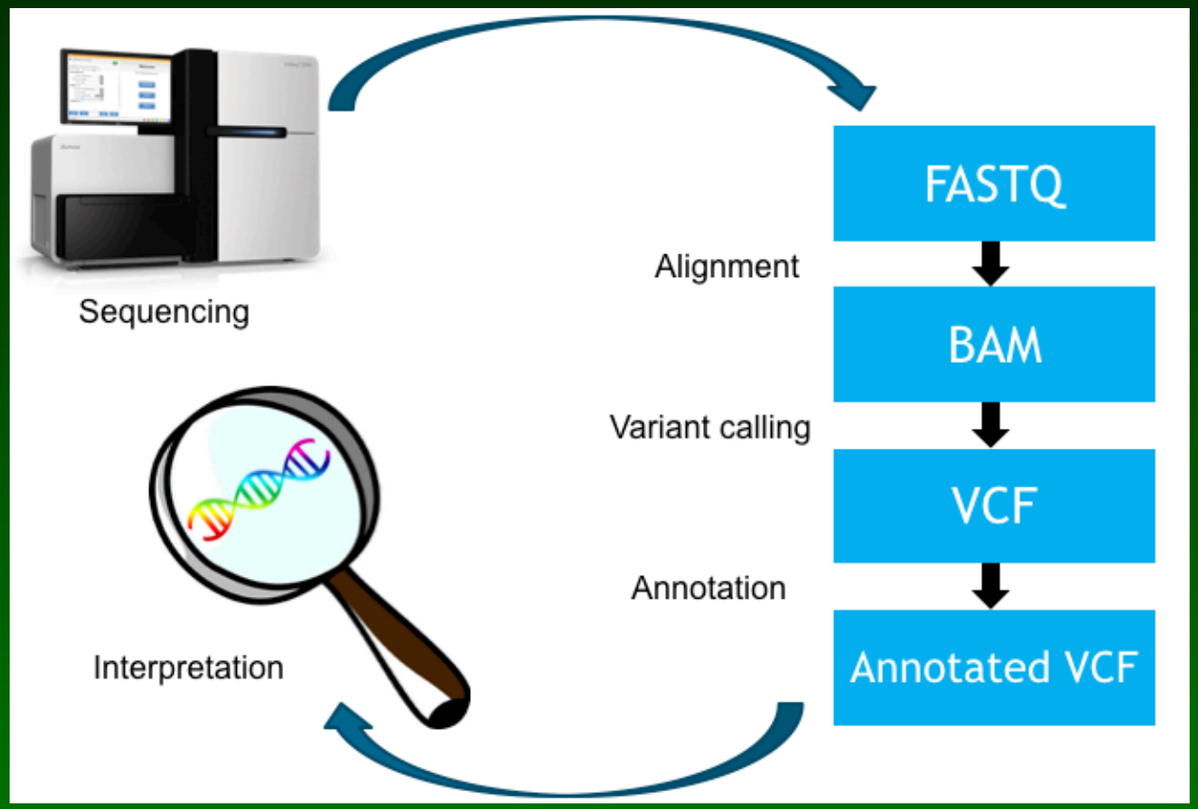


image from Christian Gilissen

Our first ever WES

- capture exome X-chromosome only
- X-linked disease
  TOD (Terminal Osseus Dysplasia)
- used pipeline with standard settings
  threshold for calling variants
    number of reads before variant is called (10 or more)
    range of accepted allele frequency

No possible causative variant remaining

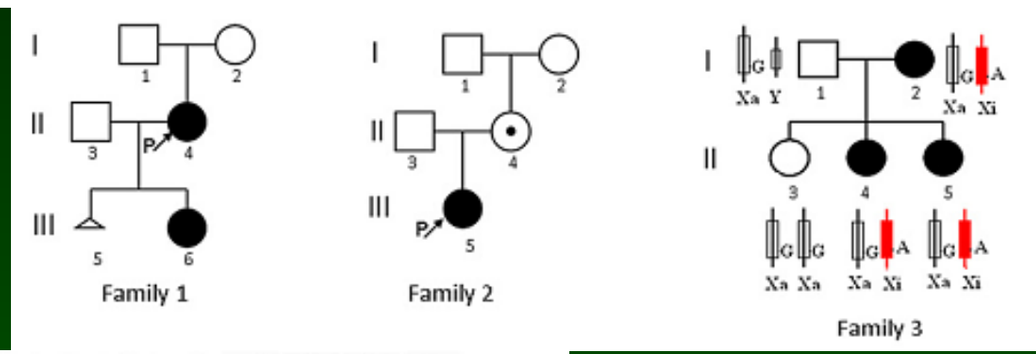
X-linked TOD

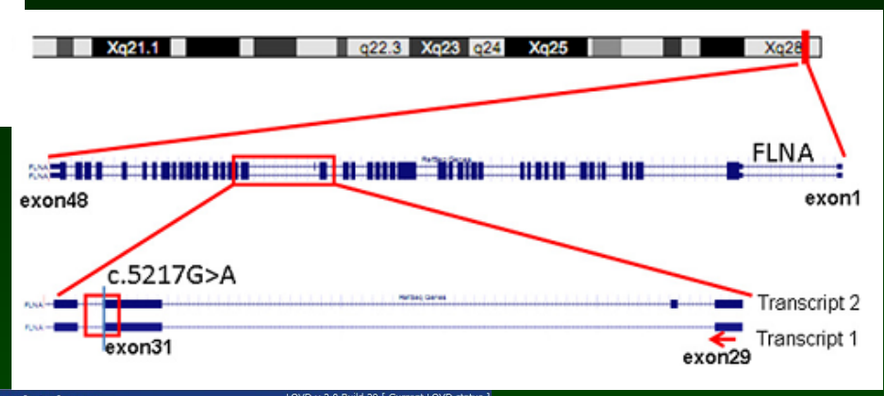
The American Journal of Human Genetics *87*, 146–153, July 9, 2010

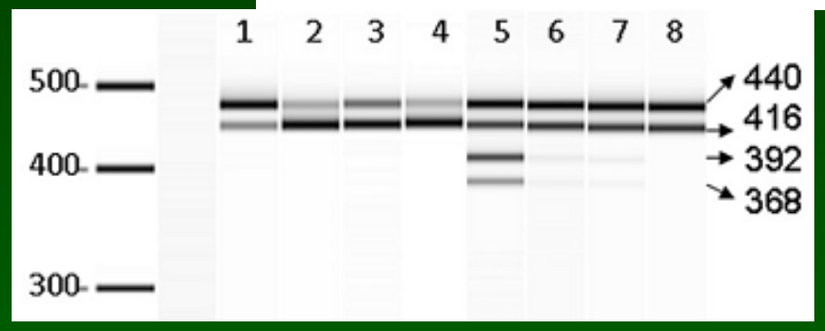
REPORT

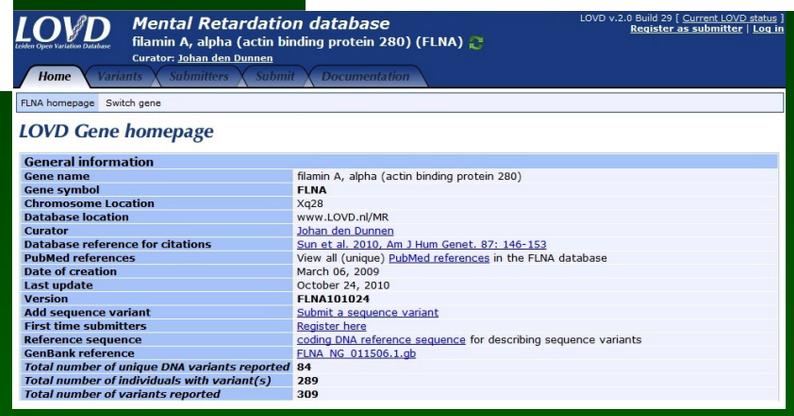
Terminal Osseous Dysplasia Is Caused by a Single Recurrent Mutation in the *FLNA* Gene

Yu Sun,[1,11] Rowida Almomani,[1,11] Emmelien Aten,[1] Jacopo Celli,[1] Jaap van der Heijden,[1] Hanka Venselaar,[2] Stephen P. Robertson,[3] Anna Baroncini,[4] Brunella Franco,[5,6] Lina Basel-Vanagaite,[7] Emiko Horii,[8] Ricardo Drut,[9] Yavuz Ariyurek,[1,10] Johan T. den Dunnen,[1,10] and Martijn H. Breuning[1,*]









General information	
Gene name	filamin A, alpha (actin binding protein 280)
Gene symbol	**FLNA**
Chromosome Location	Xq28
Database location	www.LOVD.nl/MR
Curator	Johan den Dunnen
Database reference for citations	Sun et al. 2010, Am J Hum Genet. 87: 146-153
PubMed references	View all (unique) PubMed references in the FLNA database
Date of creation	March 06, 2009
Last update	October 24, 2010
Version	**FLNA101024**
Add sequence variant	Submit a sequence variant
First time submitters	Register here
Reference sequence	coding DNA reference sequence for describing sequence variants
GenBank reference	FLNA_NG_011506.1.gb
Total number of unique DNA variants reported	**84**
Total number of individuals with variant(s)	**289**
Total number of variants reported	**309**

threshold to call variants set to 10 or more reads

we missed 9 reads, all variant

on X-chromosome (male) coverage is half

NGS workflow

- **what can go wrong ??**

...suggestions please

image from www.caracaschronicles.com

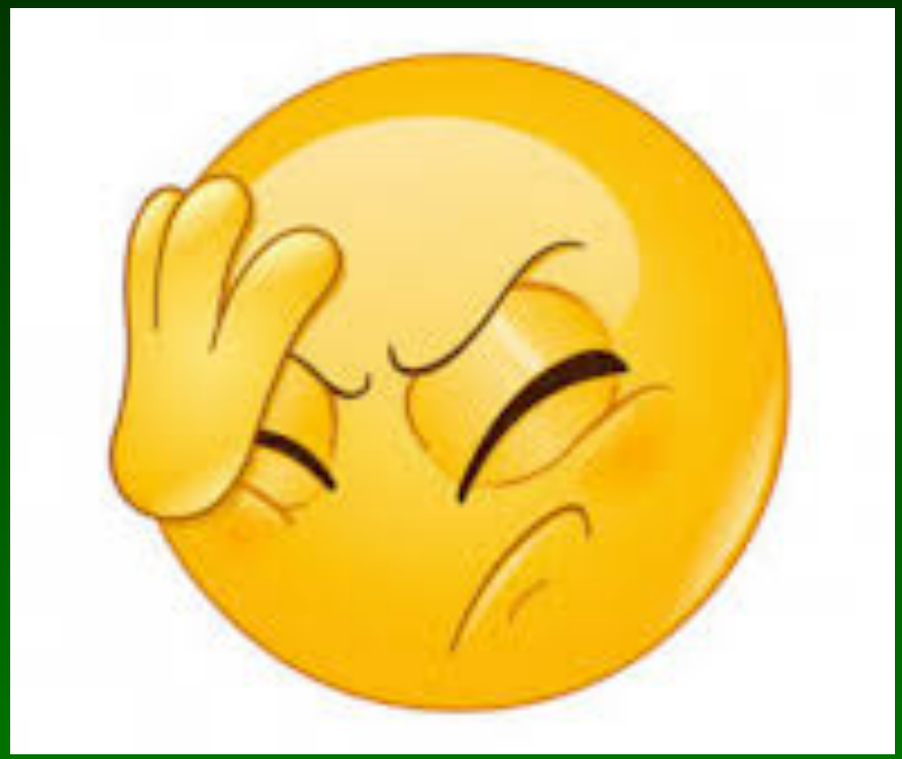


image from coinspice.io

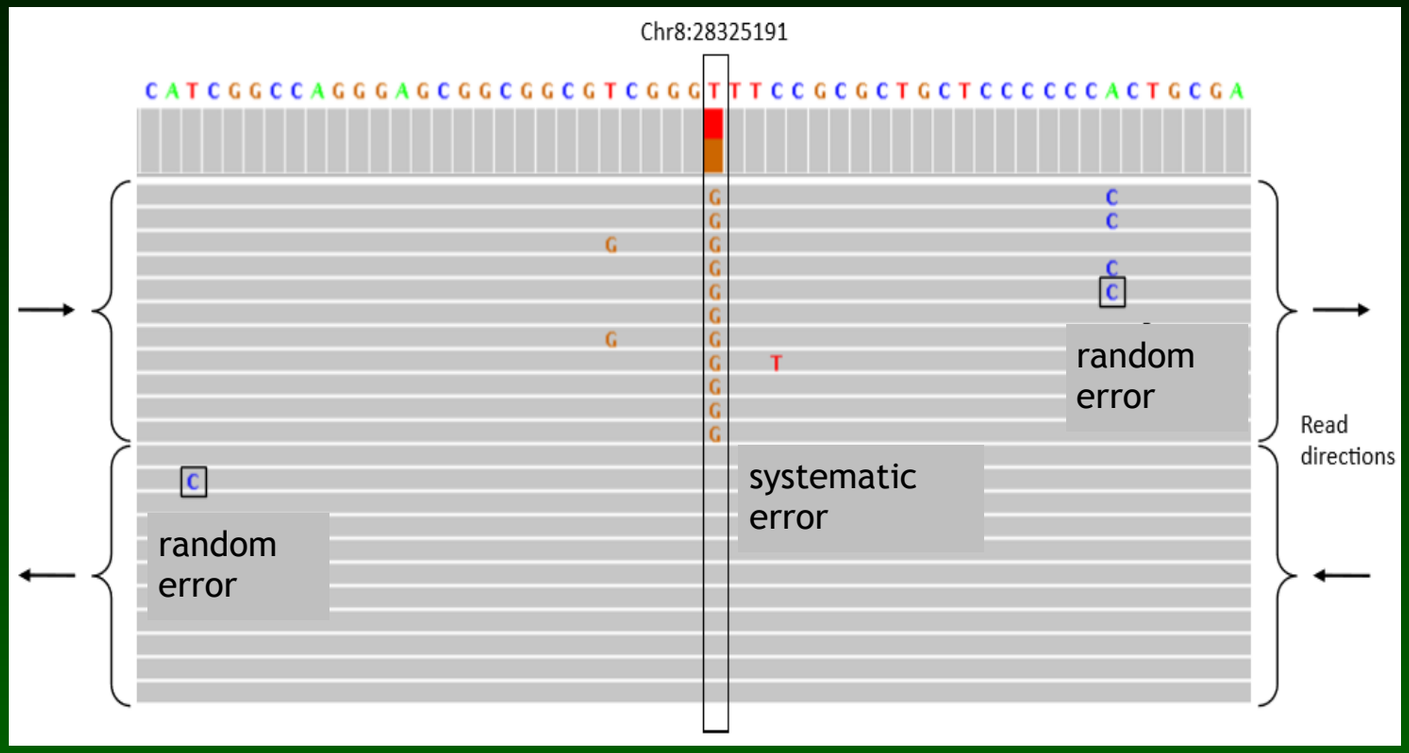
Sampling

- sampling
- sample preparation
- sequencing

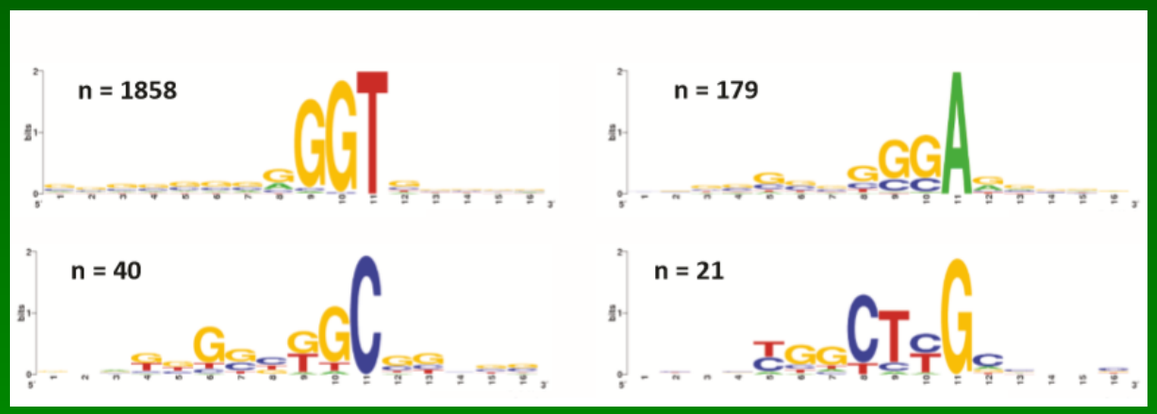
- **samples swaps will occur**
  at every step

- **build in checks**
  female / male
  SNV-assay + SNV calls from NGS
  process every sample twice
  check Mendelian inheritance

Sytematic errors

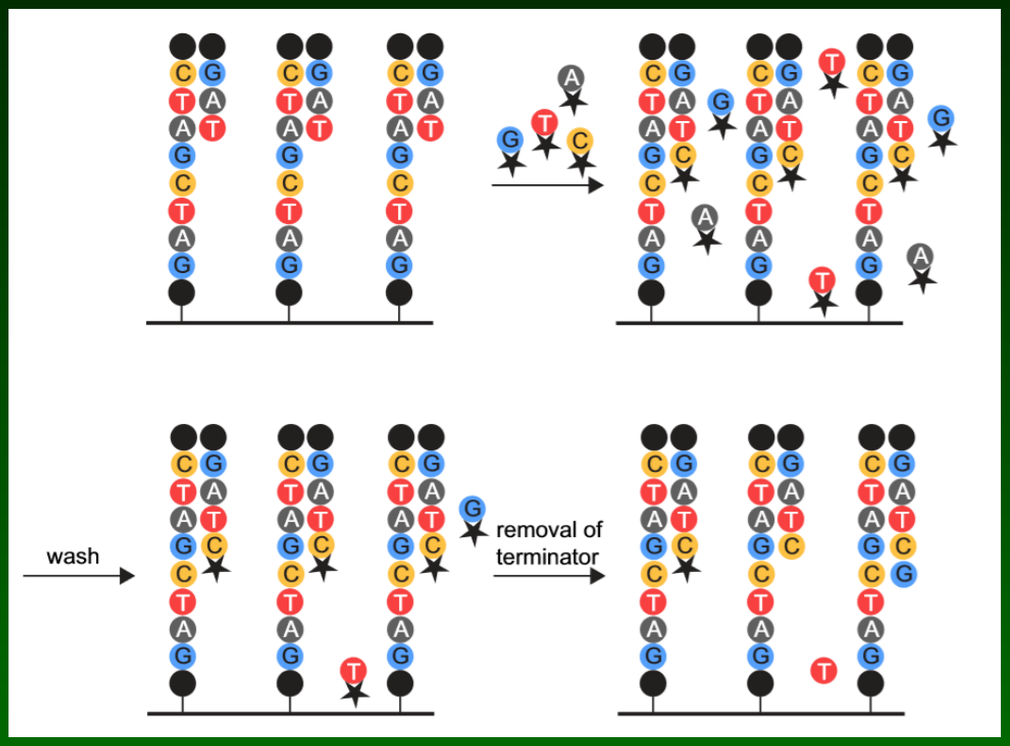


images from Pfeifer (2018)



sequence around systematic errors

origin: phasing effect



pre-phasing of strand

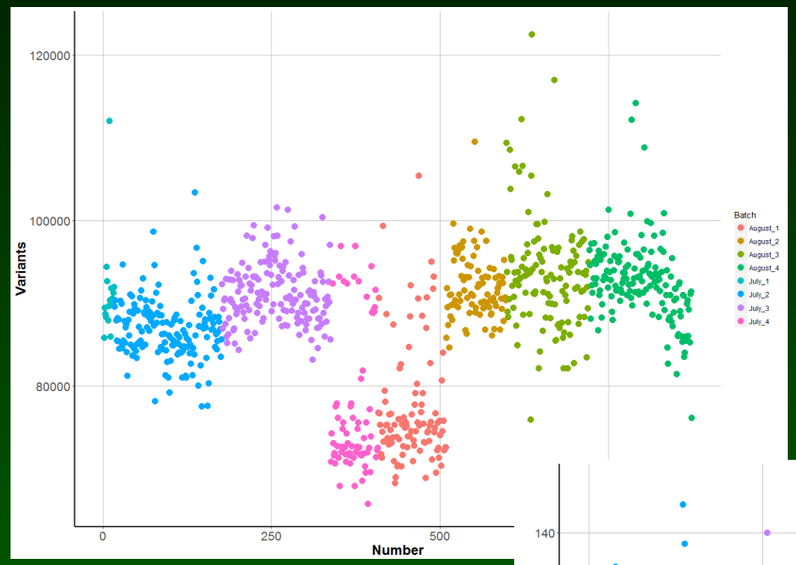
mono-nucleotide stretch problem

QC

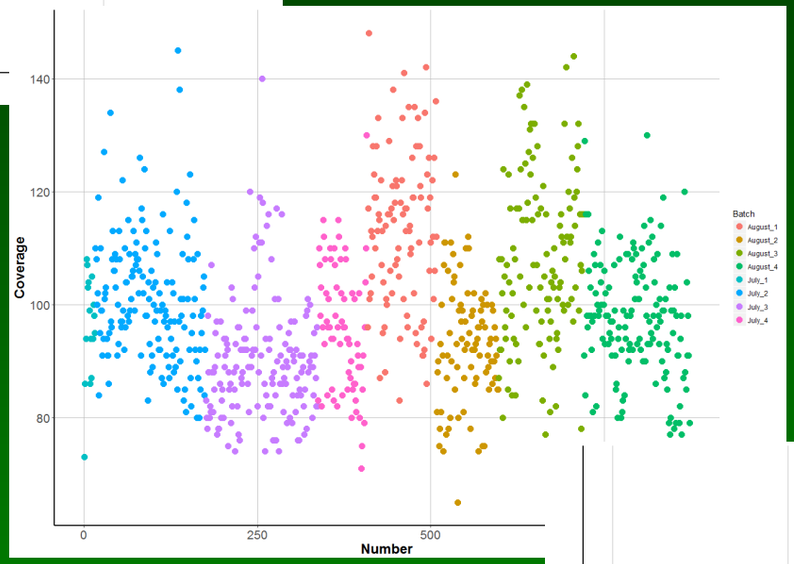
- options ?

QC

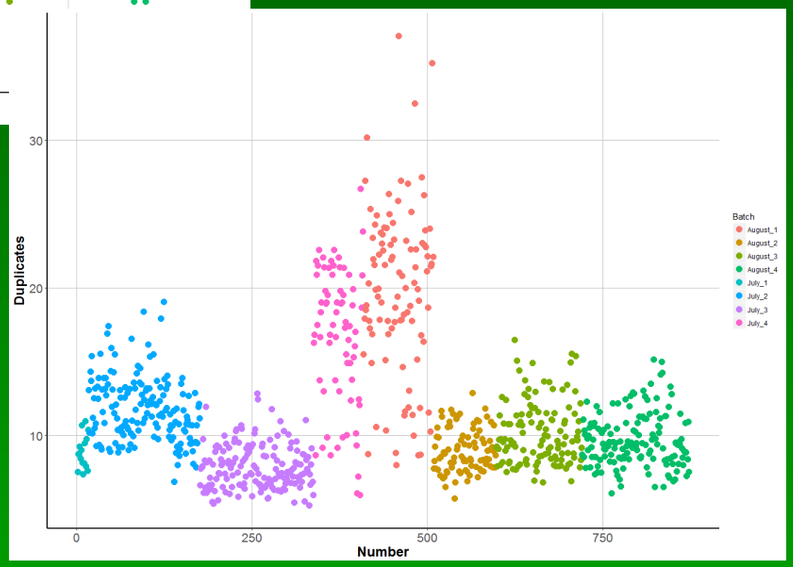
variants called per exome



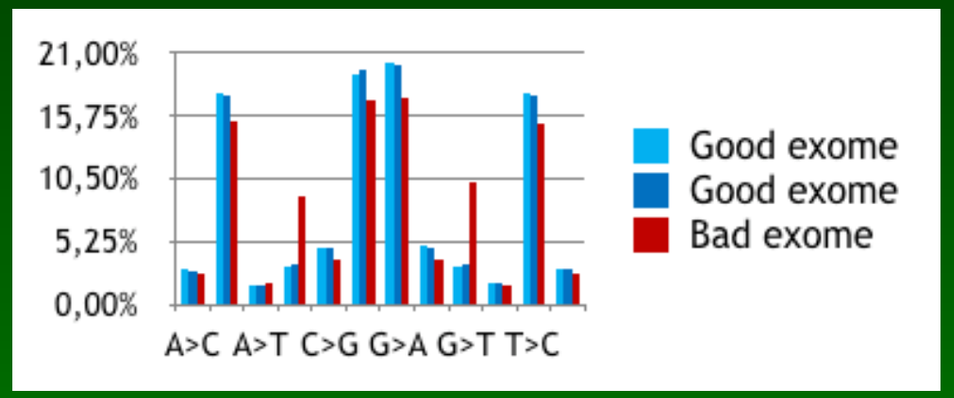
number of duplicate reads



exome coverage



% mapped reads
% reads on target
transversion / transition rate
contamination other organisms
compare with all previous analysis



© images from Christian Gilissen

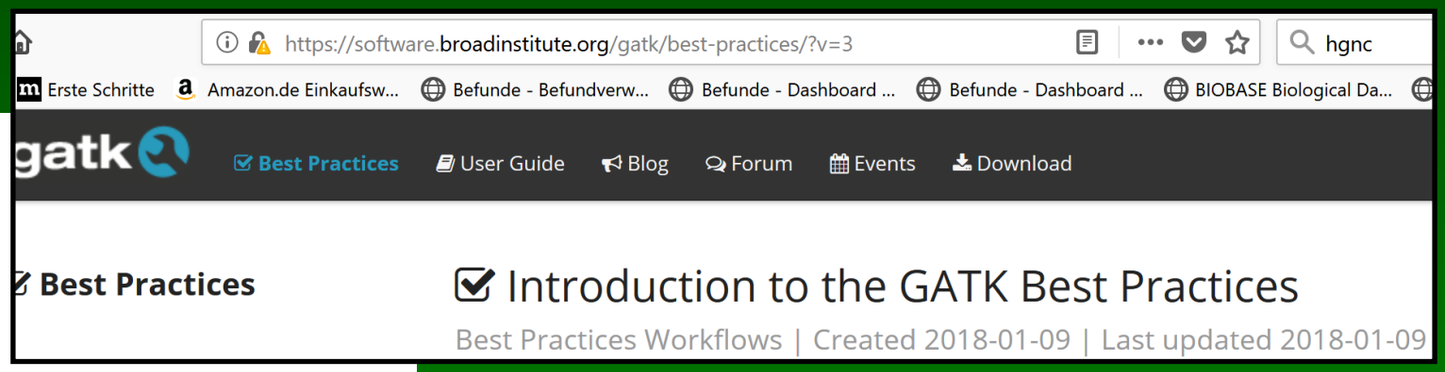
Human and Clinical Genetics

© JT den Dunnen

Updated pipeline

New pipeline / component:

first re-run standard controls and compare results



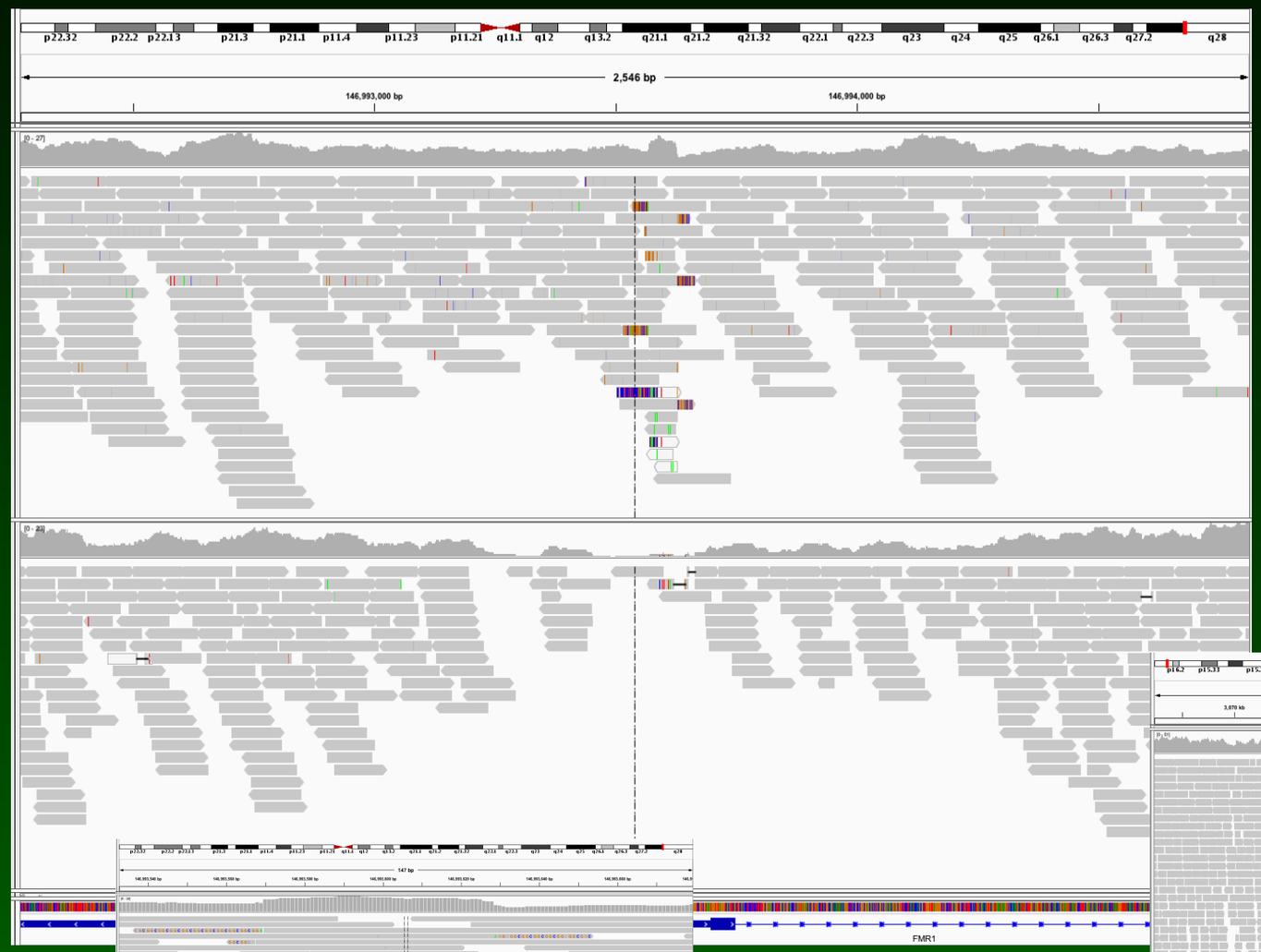
Contents

1. What are the GATK Best Practices?
2. Analysis phases
3. Experimental designs
4. Workflow scripts provided as reference implementations
5. Scope and limitations
6. What is *not* GATK Best Practices?
7. Beware legacy scripts

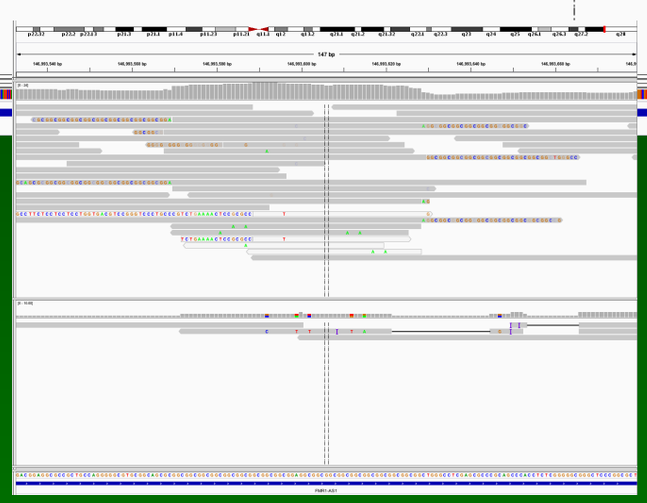
Uneven coverage - CG%

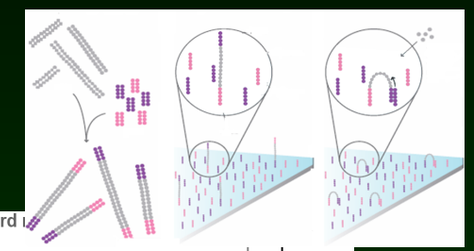
FMR1

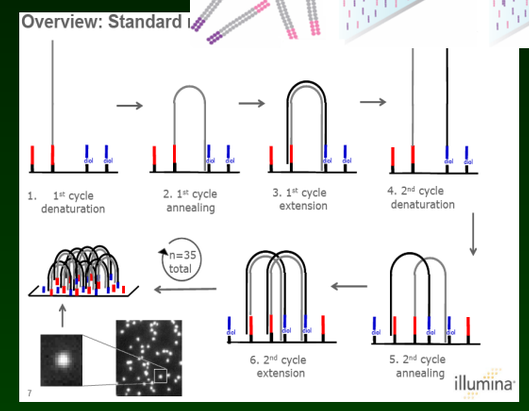
without PCR

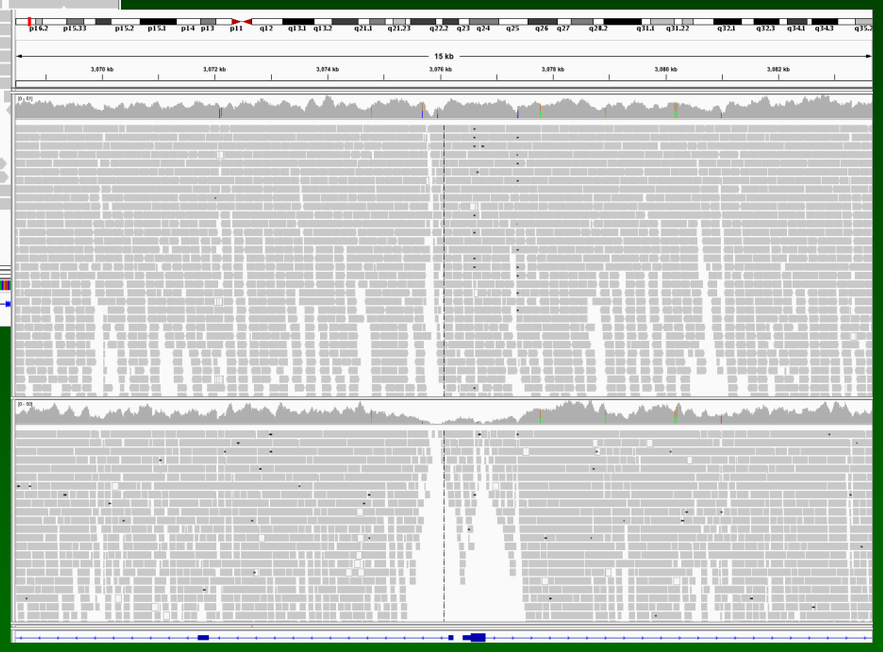


with PCR





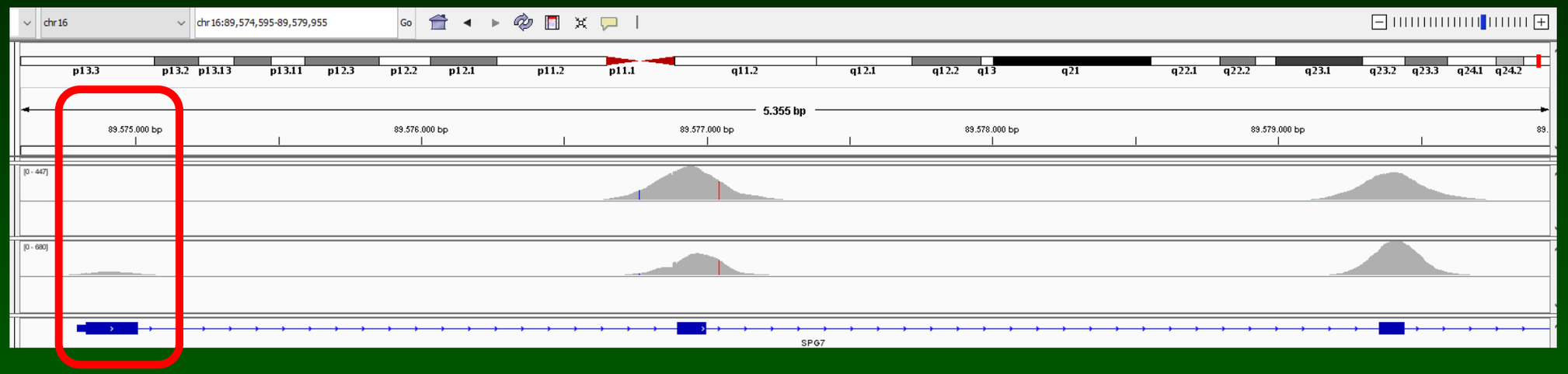


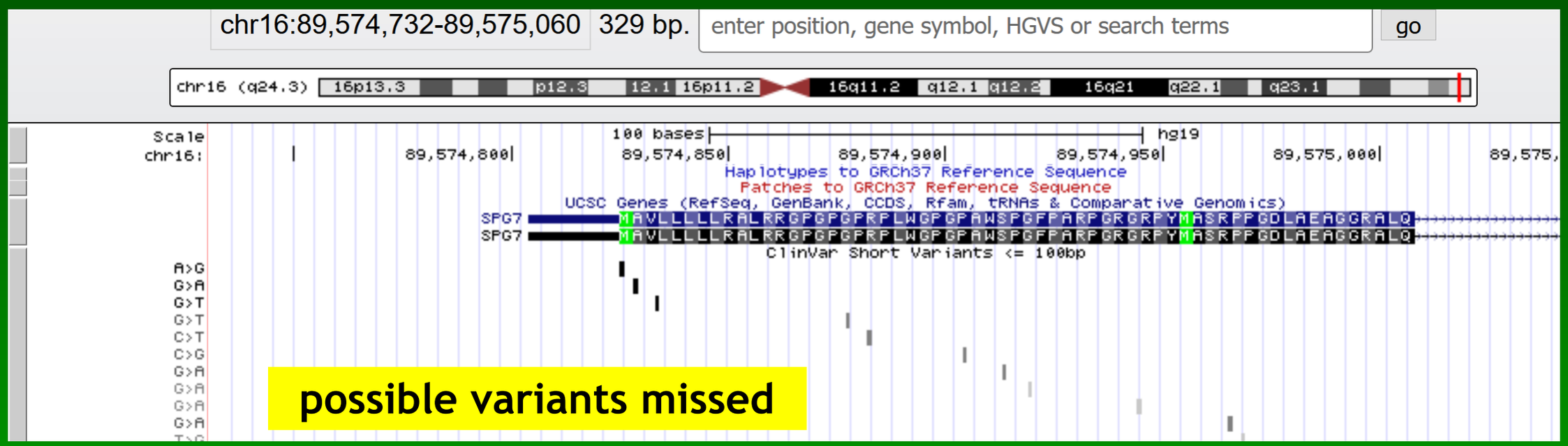


HTT

Uneven coverage

sample preparation
gene panel / exome capture





possible variants missed

Different tools

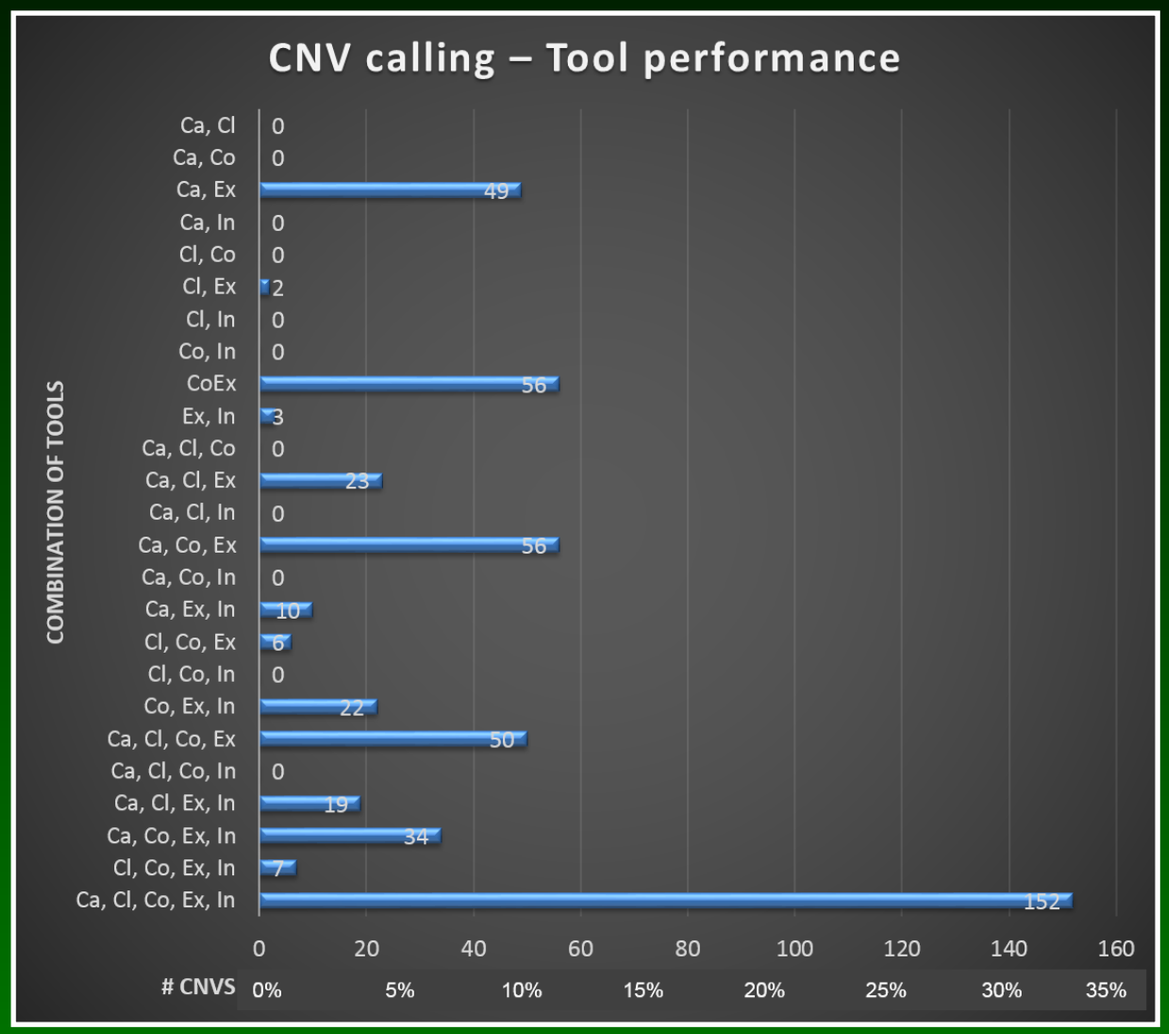


image from Anna Benet-Pages

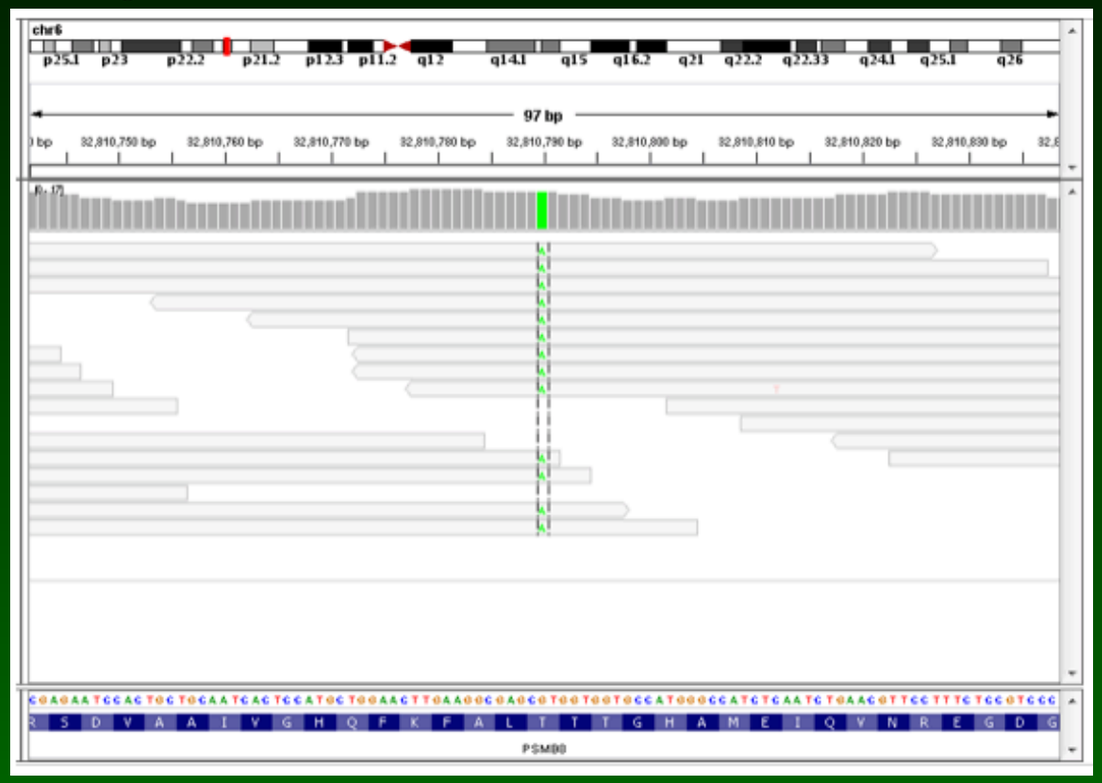
…different results

Tool	Mean Calls per Sample
Canoes	~ 1.4
Clamms	~ 5.1
Codex	~ 2.6
ExomeDepth	~ 7.7
Inhouse	~ 4.4

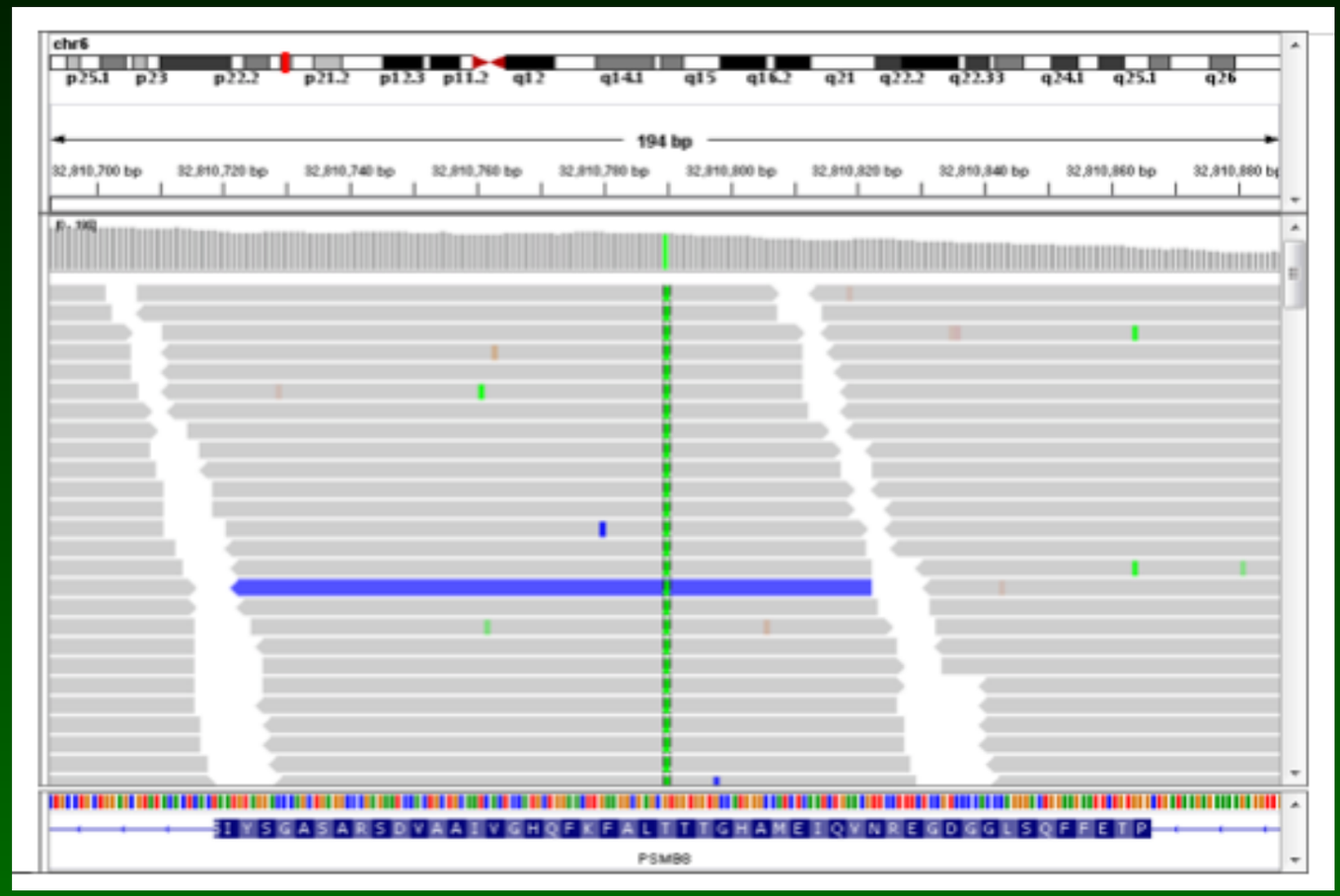
esp. for CNV detection given coverage is variable

Mapping reads

images from Christian Gilissen



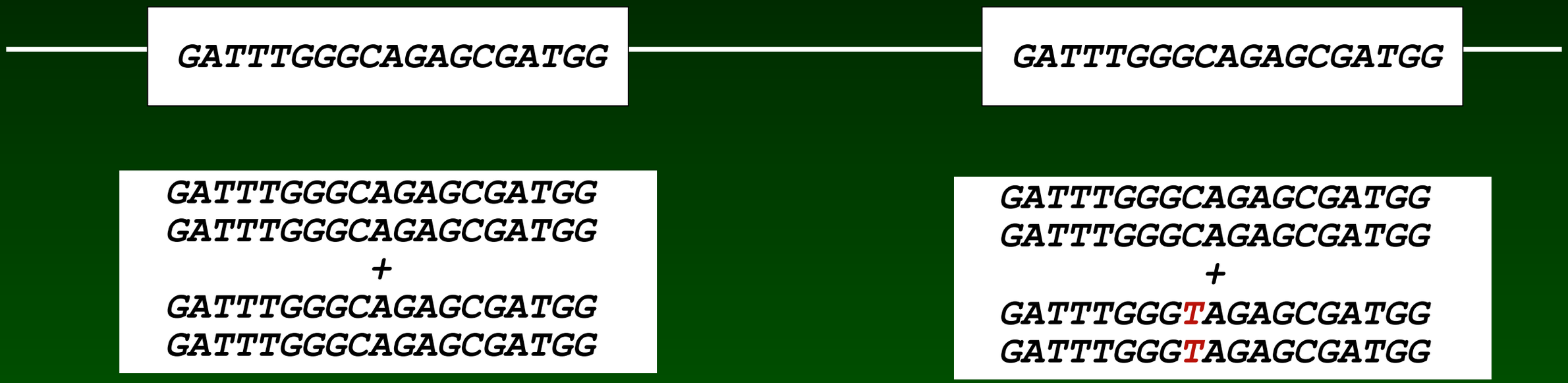
mapped to reference genome
+ unmapped contigs
+ alternate haplotypes



mapped to reference genome
+ unmapped contigs
(no alternate haploptypes)

Mapping

true situation



0% variant ***50% variant***

(e.g. globin genes)

Non-unique$_2$

goes to dust bin

no coverage

"deletion"

no coverage

"deletion"

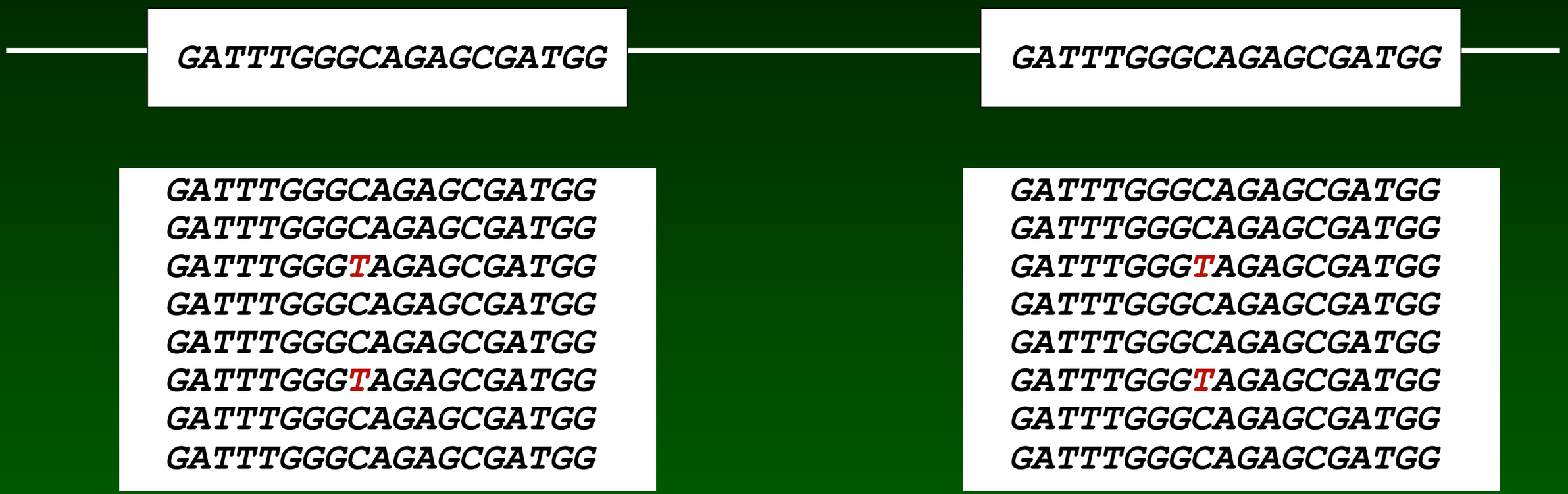
Non-unique$_3$

map to first position

GATTTGGGCAGAGCGATGG	GATTTGGGCAGAGCGATGG
GATTTGGGCAGAGCGATGG GATTTGGGCAGAGCGATGG GATTTGGGTAGAGCGATGG GATTTGGGCAGAGCGATGG GATTTGGGCAGAGCGATGG GATTTGGGTAGAGCGATGG GATTTGGGCAGAGCGATGG GATTTGGGCAGAGCGATGG	
25% variant	***(no data)***

Non-unique$_2$

map to all positions



25% variant

25% variant

variant missed / found twice

Non-unique$_4$

probabalistic mapping

GATTTGGGCAGAGCGATGG

GATTTGGGCAGAGCGATGG

GATTTGGGCAGAGCGATGG
GATTTGGGTAGAGCGATGG
GATTTGGGCAGAGCGATGG
GATTTGGGCAGAGCGATGG

25% variant

GATTTGGGCAGAGCGATGG
GATTTGGGTAGAGCGATGG
GATTTGGGTAGAGCGATGG
GATTTGGGCAGAGCGATGG

50% variant

0% variant

or

or

GATTTGGGCAGAGCGATGG
GATTTGGGCAGAGCGATGG
GATTTGGGCAGAGCGATGG
GATTTGGGTAGAGCGATGG

25% variant

GATTTGGGCAGAGCGATGG
GATTTGGGCAGAGCGATGG
GATTTGGGCAGAGCGATGG
GATTTGGGCAGAGCGATGG

0% variant

50% variant

Gene conversion

Pseudoxanthoma elasticum: c.4182del (p.Lys1394Asnfs*9)

exon 9 cooverage 0.5 **ABCC6** **ABCC6P1** exon 9 coverage 1.5

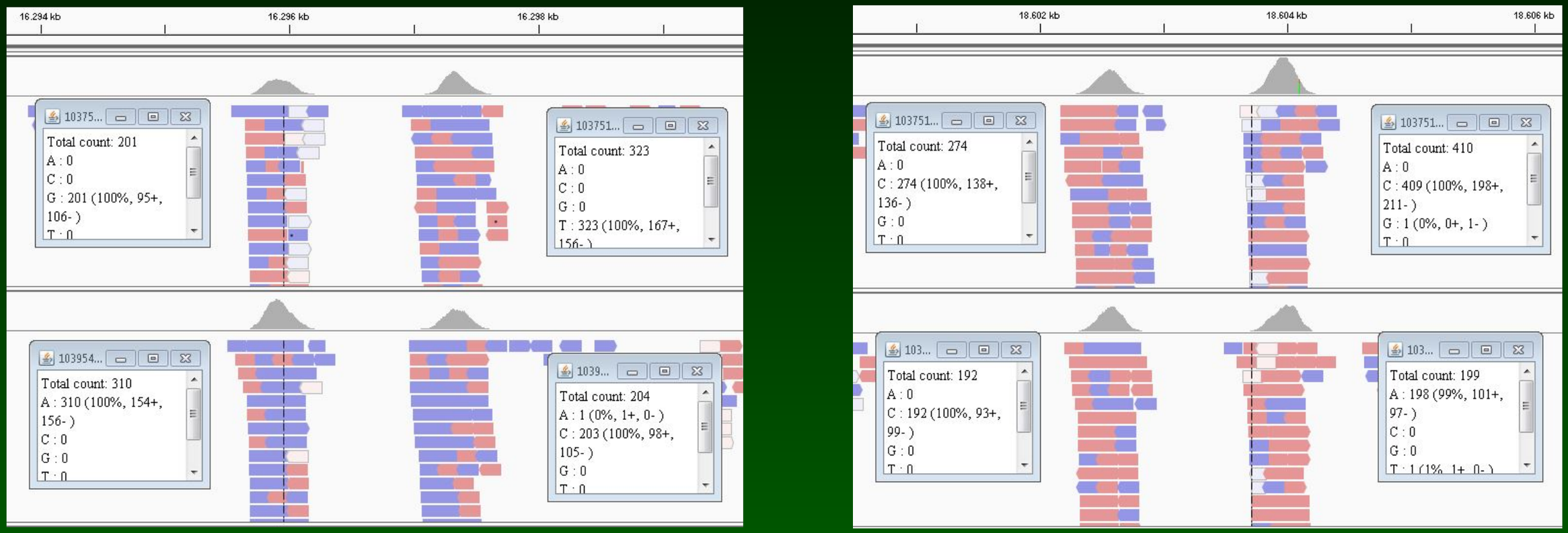


image from Anna Benet-Pages

gene conversion causes reads to map to ABCC6P1

no variant in ABCC6

no variant pseudogene

coverage differs !!

Homologous sequences

VWF gene causes "*von Willebrand Syndrome*"

mechanism: SNVs (~90%), CNVs (~10%), 6-335 bp gene conversions

VWF - von Willebrand factor gene
chr12:6,118,707-6,137,253

VWFP1 - von Willebrand factor pseudogene 1
chr22:17,170,925-17,185,367

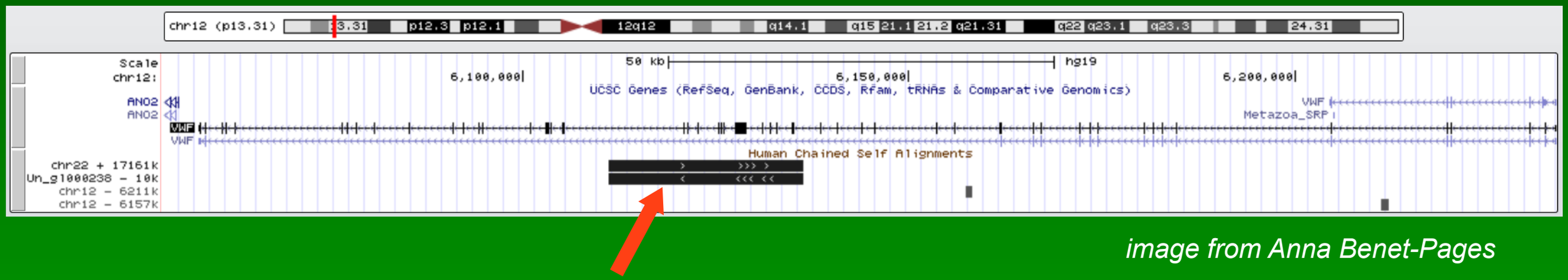


image from Anna Benet-Pages

Calling variants

whole exome sequencing

call variants in targeted regions

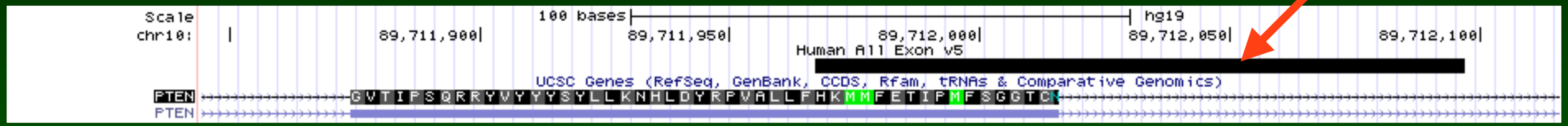


image from Christian Gilissen

probe = targeted but you capture more

call variants in targeted regions +/- … nucleotides

Calling thresholds

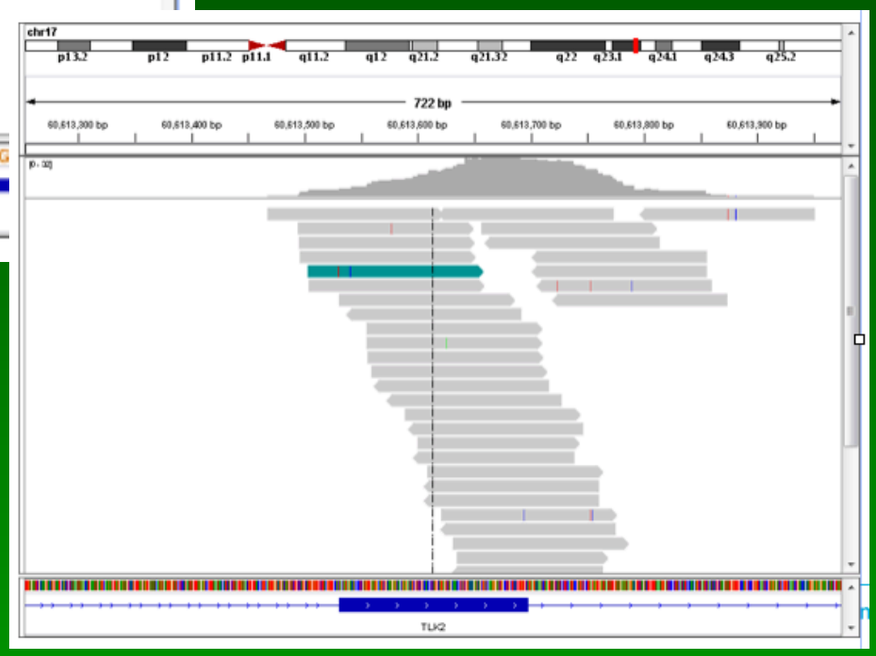
image from Christian Gilissen

Thresholds are essential

too low = many false positive
too high = miss true variants

a delicate balance

NOTE: X/Y coverage in males



De novo variants

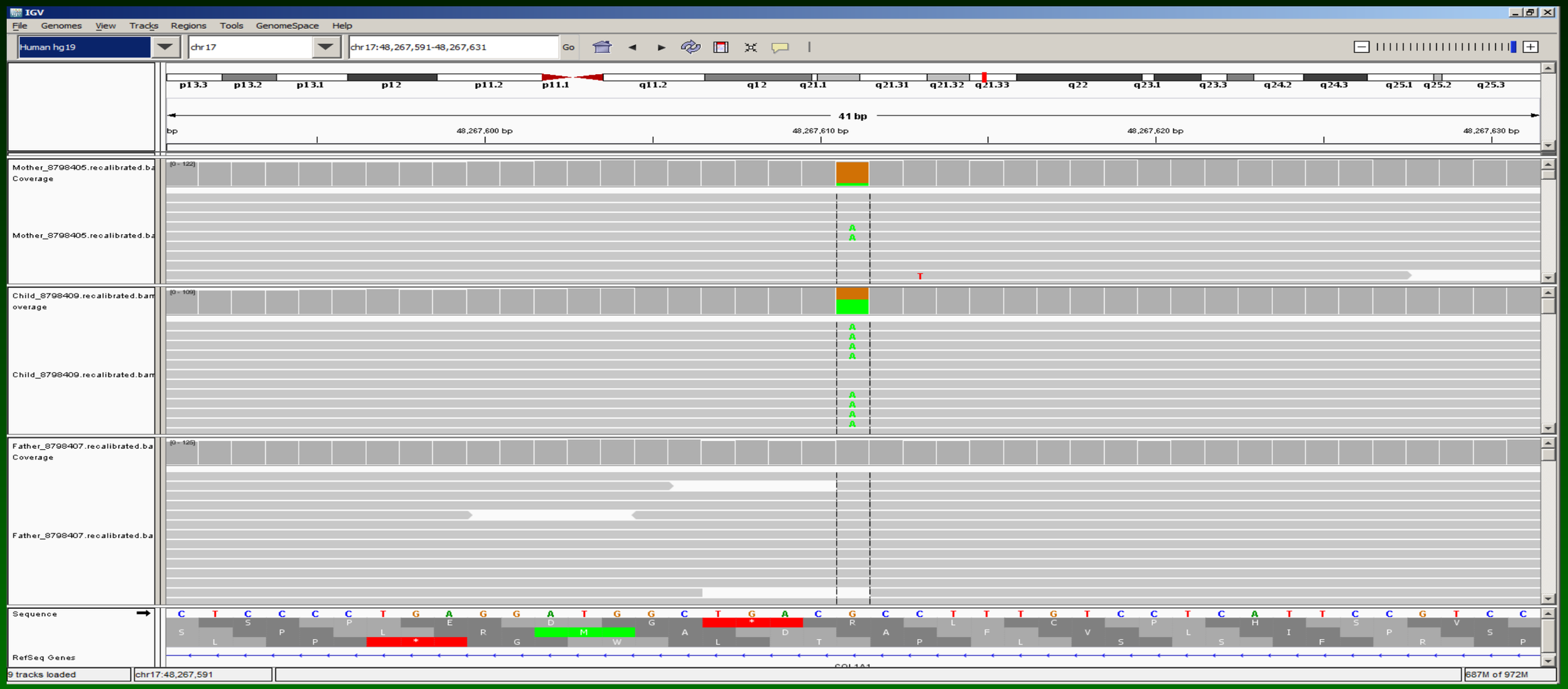


image from Gijs Santen

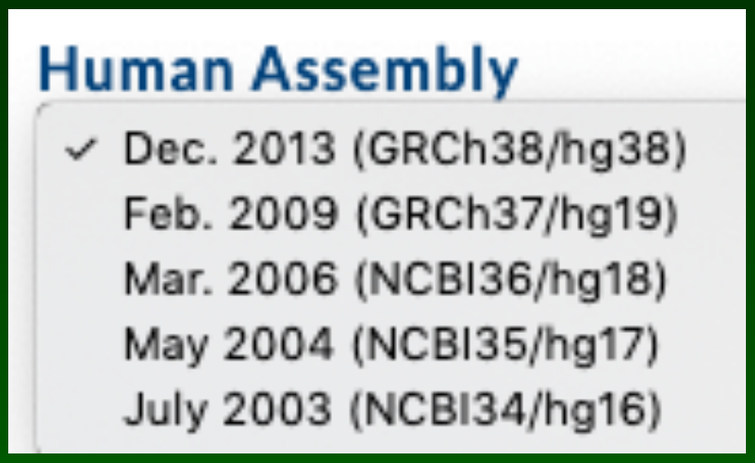
variant missed as de novo

why? ***no variant read allowed in parent***

Reference genome

Which genome build do you use?

From which year is it ?



Is the genome complete ?

XYLT1

- **known disease gene**
  Desbuquois dysplasia type 2 (DBQD2)
  many patients without deleterious variants

- **detailed analysis negative**
  Sanger sequencing, WES, WGS

- **XYLT1 methylated in patients**
  allele not expressed

GGC Repeat Expansion and Exon 1 Methylation of *XYLT1* Is a Common Pathogenic Variant in Baratela-Scott Syndrome

American Journal of Human Genetics *104*, 35–44

Amy J. LaCroix,[1,11] Deborah Stabley,[2,11] Rebecca Sahraoui,[2,3] Margaret P. Adam,[1,4] Michele Mehaffey,[1] Kelly Kernan,[1] Candace T. Myers,[5] Carrie Fagerstrom,[6] George Anadiotis,[6] Yassmine M. Akkari,[6] Katherine M. Robbins,[2] Karen W. Gripp,[2] Wagner A.R. Baratela,[7,8] Michael B. Bober,[7] Angela L. Duker,[7] Dan Doherty,[1,4] Jennifer C. Dempsey,[1] Daniel G. Miller,[1] Martin Kircher,[9] Michael J. Bamshad,[1,4] Deborah A. Nickerson,[4,9] University of Washington Center for Mendelian Genomics, Heather C. Mefford,[1,4,12,*] and Katia Sol-Church[2,10,12,*]

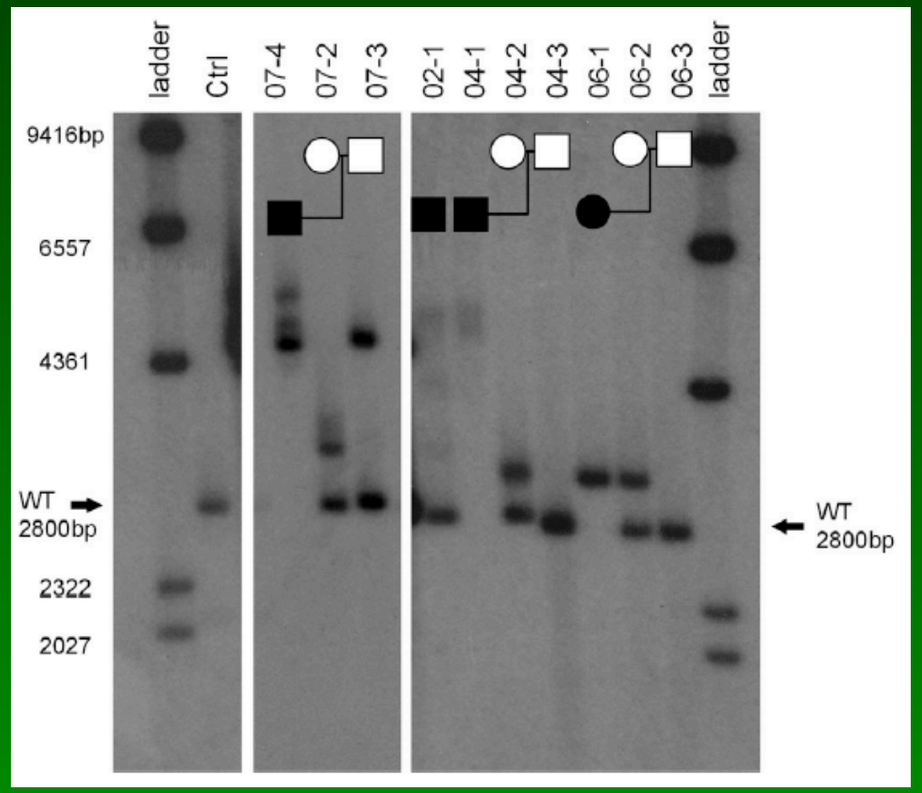
XYLT1

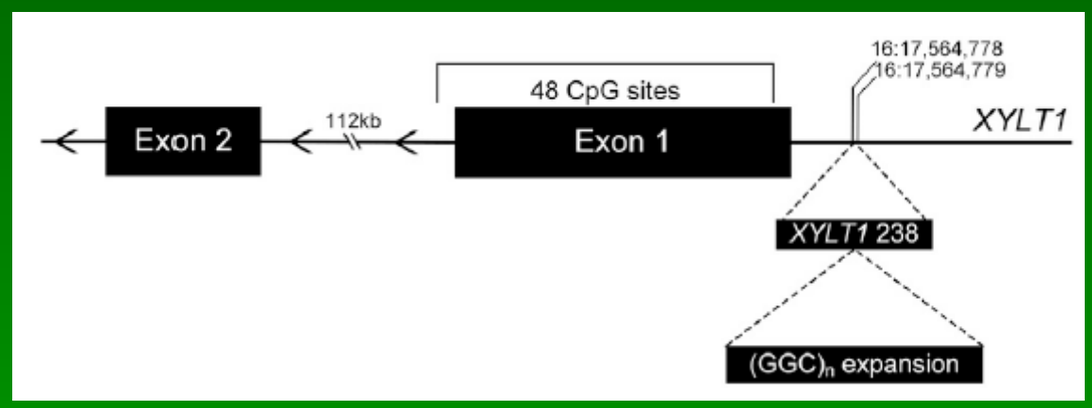
GGC Repeat Expansion and Exon 1 Methylation of *XYLT1* Is a Common Pathogenic Variant in Baratela-Scott Syndrome

American Journal of Human Genetics *104*, 35–44

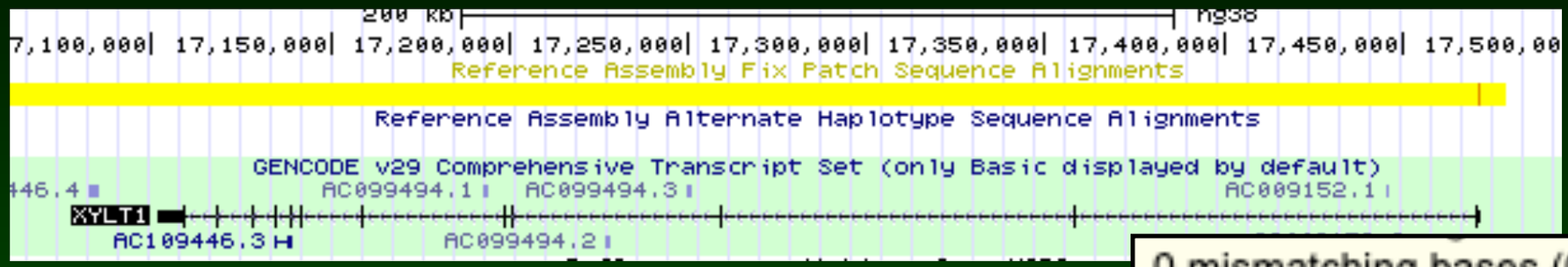
Amy J. LaCroix,[1,11] Deborah Stabley,[2,11] Rebecca Sahraoui,[2,3] Margaret P. Adam,[1,4] Michele Mehaffey,[1] Kelly Kernan,[1] Candace T. Myers,[5] Carrie Fagerstrom,[6] George Anadiotis,[6] Yassmine M. Akkari,[6] Katherine M. Robbins,[2] Karen W. Gripp,[2] Wagner A.R. Baratela,[7,8] Michael B. Bo
Dan Doherty,[1,4] Jennifer C. Dempsey,[1] Daniel G. Miller,[1] Martin Kircher,[9] Mich
Deborah A. Nickerson,[4,9] University of Washington Center for Mendelian Genc
Heather C. Mefford,[1,4,12,*] and Katia Sol-Church[2,10,12,*]

sulfate and dermatan sulfate proteoglycans.[9] The promoter region of *XYLT1* was recently characterized and found to have 238 bp of sequence (XYLT1-238) that is not present in the reference genome (GRCh37/hg19, GRCh38/hg38) but appears to be fixed in the human population (100/100 individuals tested)[10] and is evolutionarily conserved in chimpanzee and mouse. The 238 bp region also contains a variable GGC repeat with a range of 9–20 repeats in 100 individuals tested; repeat length within this range did not

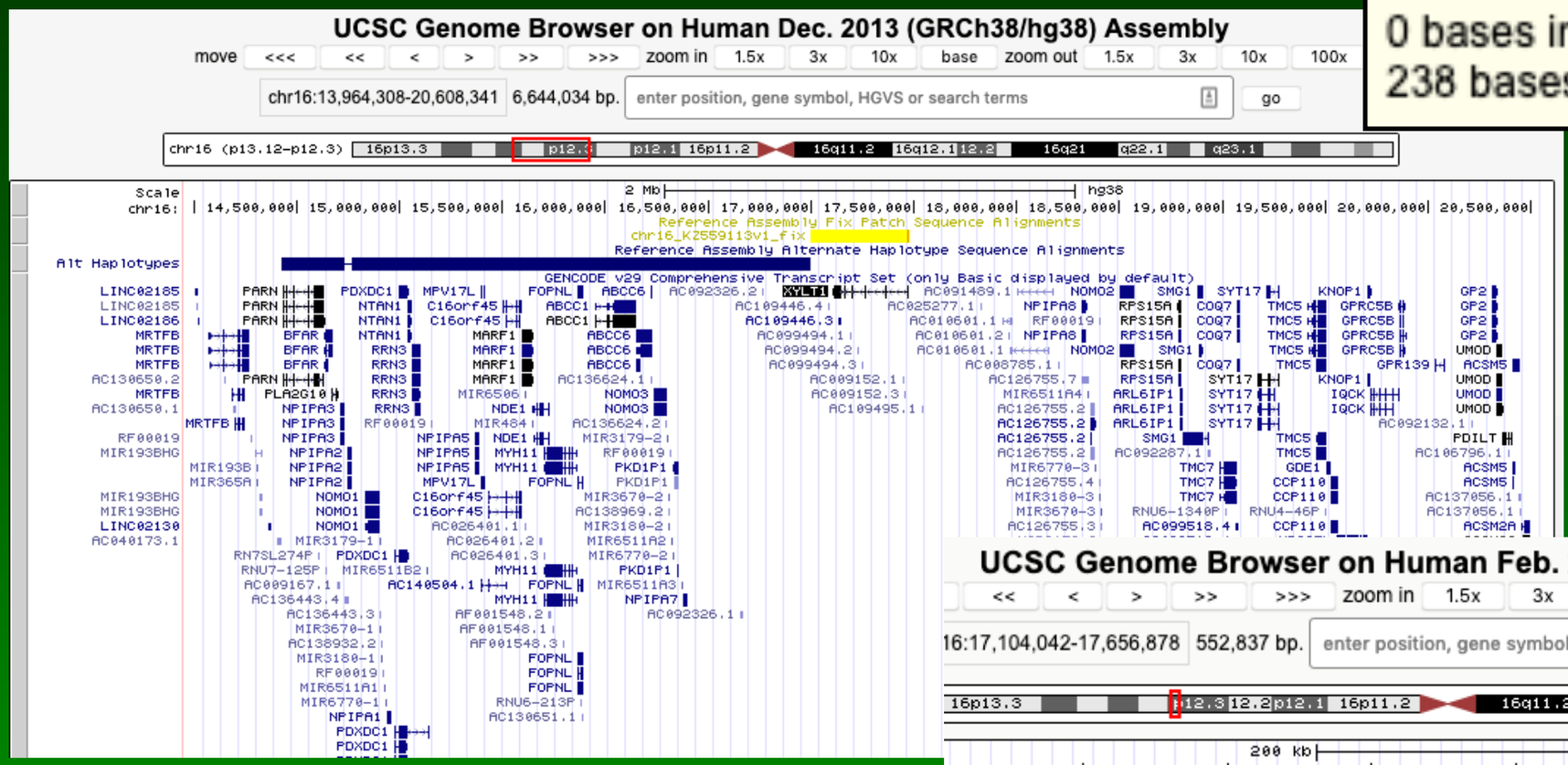


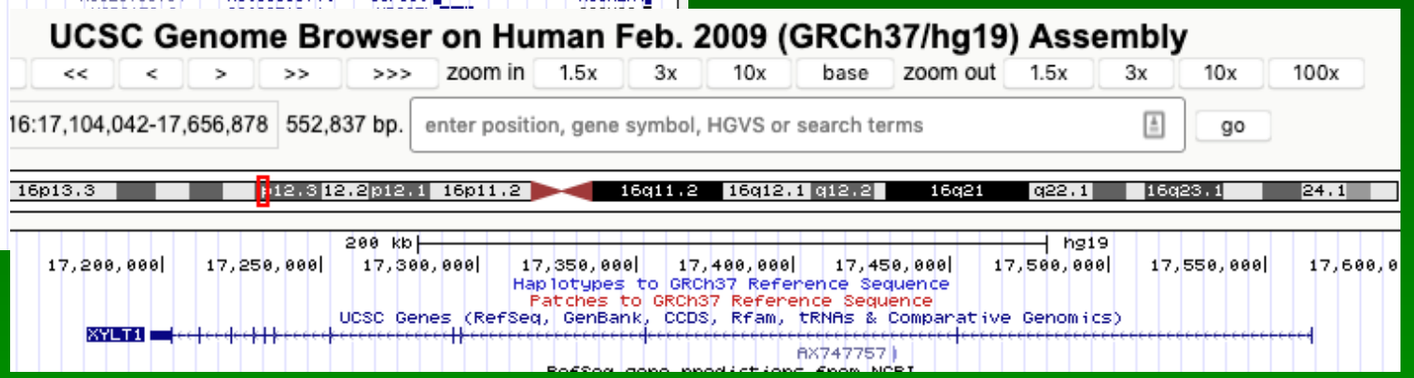


XYLT1 hg38



0 mismatching bases (0.00%)
0 N bases(0.00%)
0 bases inserted in chr16
238 bases inserted in chr16_KZ559113v1_fix





updated in hg38
(patch, alternate haplotype)

Dutch genome

Go • NL
GENOMEoftheNETHERLANDS

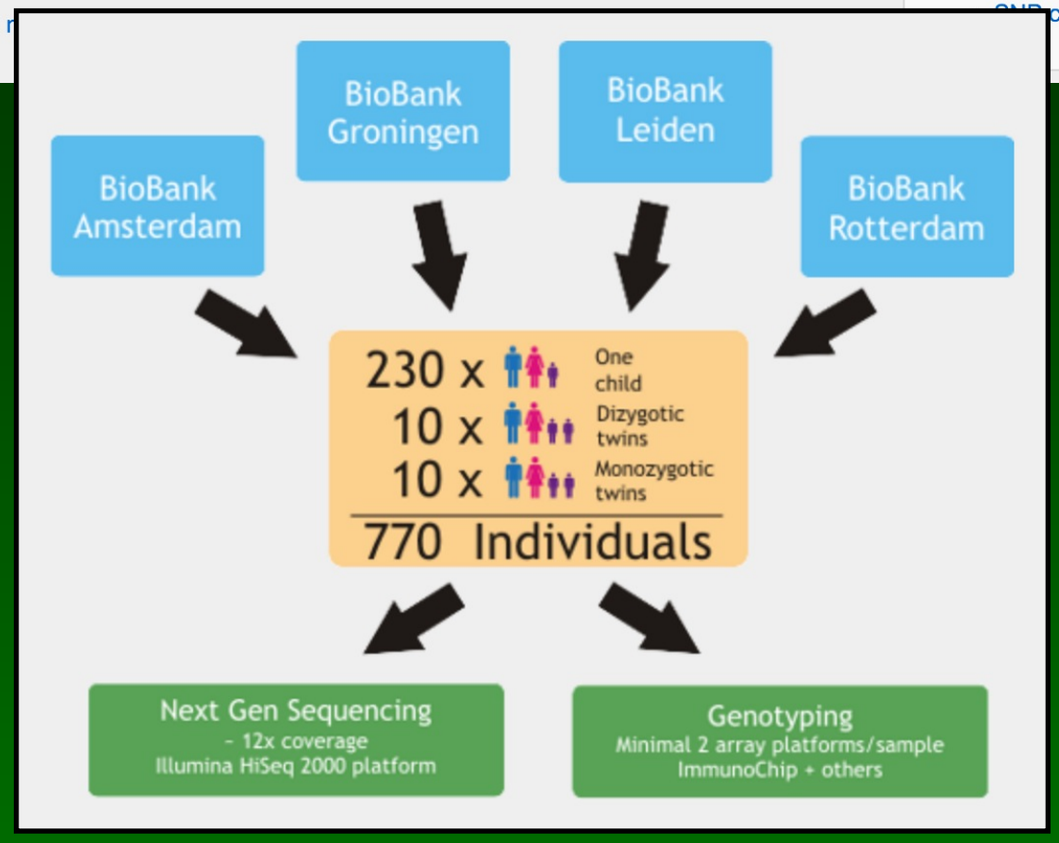
Home | About us | Access to the data | Resources | The GoNL team | Wiki

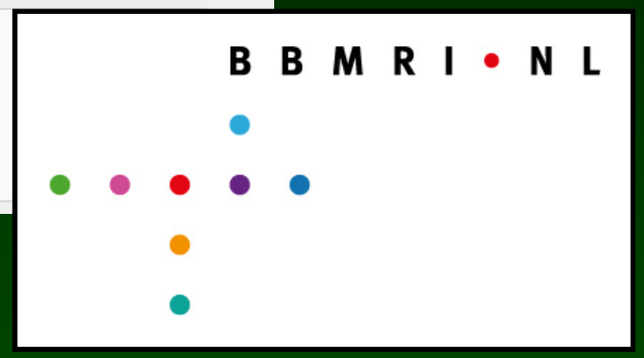
Ultra-sharp genetic group portrait of the Dutch

Posted on July 15, 2012 by r

News

SNP calling complete





used as reference ?

contains sequence not found in reference genome

Repeat expansion

- **disease gene mapped for many years**
  - ***extensive families available***
  - ***…no deleterious variants***
  - ***Sanger sequencing, WES, WGS, RNA-seq***

nature genetics
ARTICLES
https://doi.org/10.1038/s41588-018-0067-2

Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy

Hiroyuki Ishiura[1], Koichiro Doi[2], Jun Mitsui[1], …
Asao Fujiyama[3], Yasuko Toyoshima[4], Akiyoshi Ka…
Sumio Sugano[6], Wei Qu[2], Kazuki Ichikawa[2], Hide…

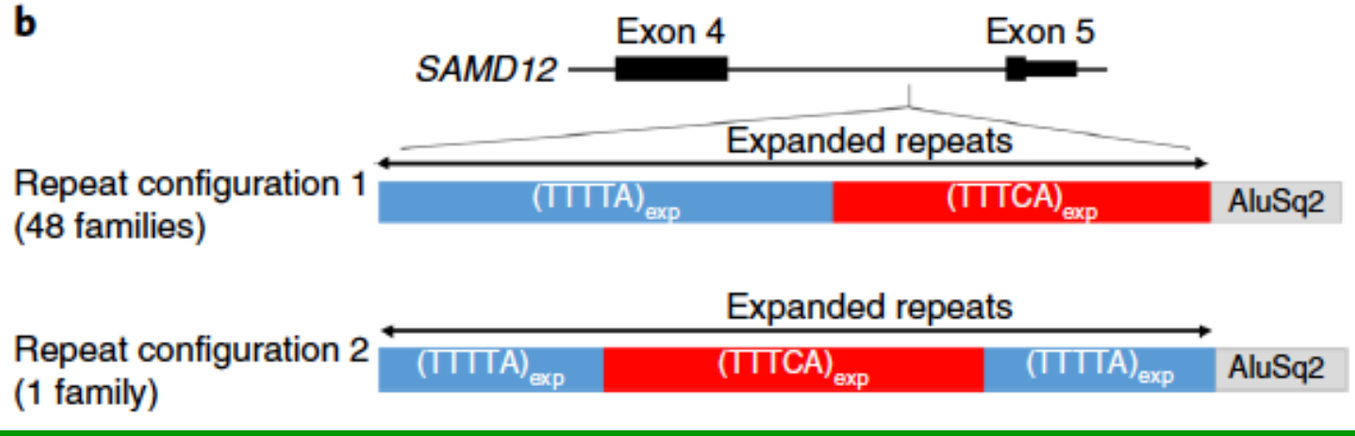
expanded sequence not in reference

image from www.dreamstime.com

Reference genome 5′-GTTCAT-$(TTTTA)_7$TTA$(TTTTA)_{13}$-TTTGA

F6906 (II-6) 5′-GTTCAT-$(TTTTA)_2$TTTTTTA$(TTTTA)_{104}$TTA$(TTTTA)_3$TTTA$(TTTTA)_{10}$$(TTTCA)_{66}$-TTGA

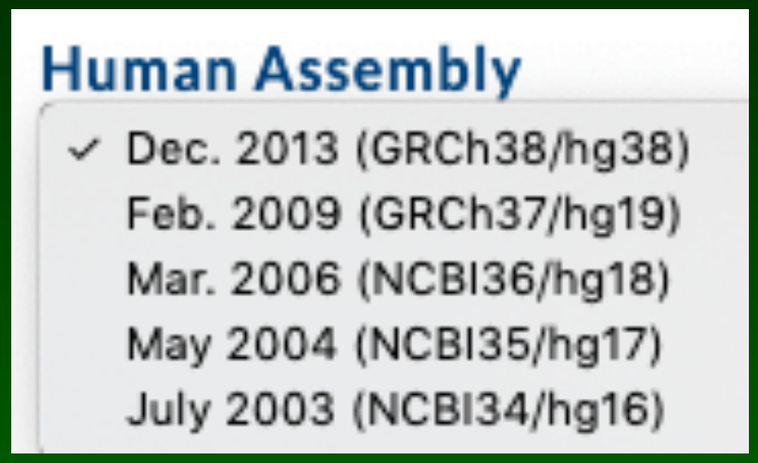
F6115 (I-2) 5′-GTTCAT-$(TTTTA)_3$TTTTTTTA$(TTTTA)_{60}$ $(TTTCA)_{61}$$(TTTTA)_{73}$TTTATTTTA-TTTGA



Reference genome

Which genome build do you use?

From which year is it ?



Is the genome complete ?

Is the transcript annotation complete ?

Gene annotation

call variants in genes / transcripts

RefSeq

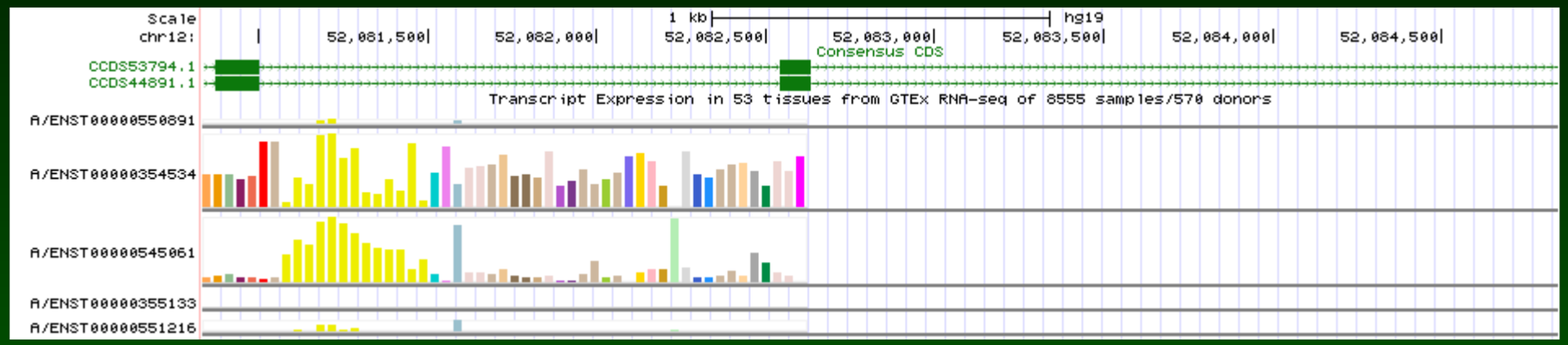
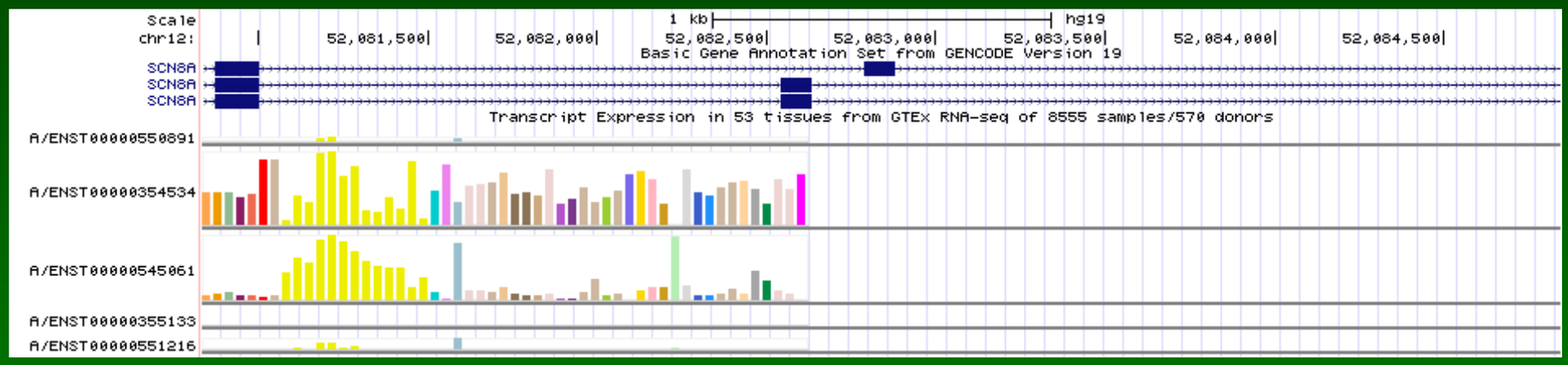


image from Christian Gilissen

...



calling variants linked to all transcripts gives an enormous list

select transcripts, but be careful

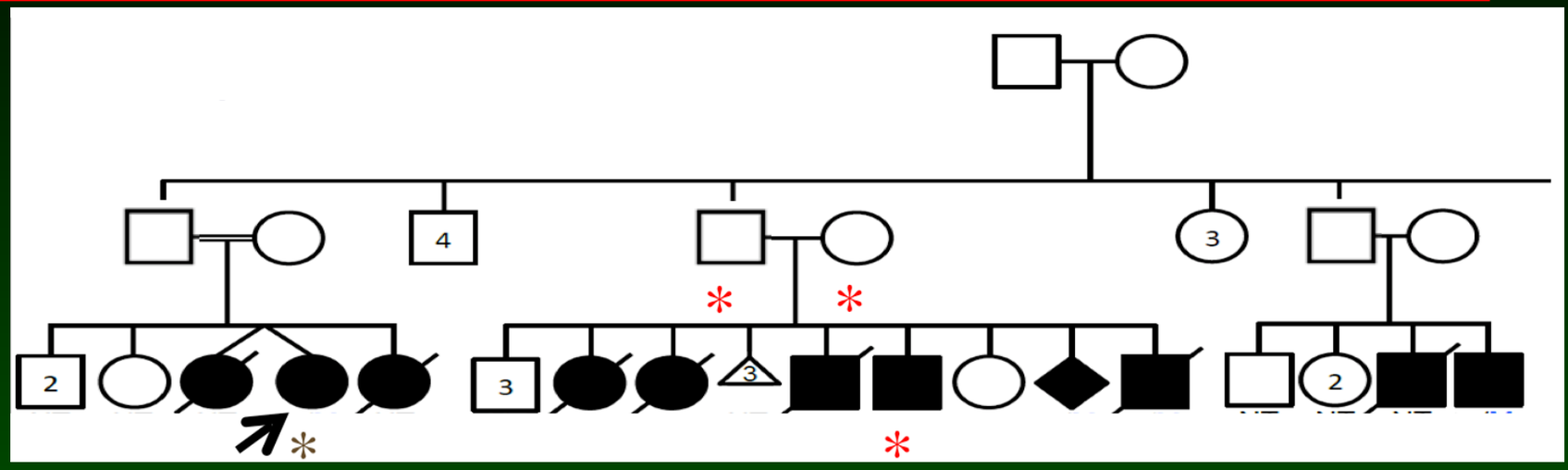
Frequently variant

Shyr (2014) BMC Medical Genomics

not likely candidates:

- **alignment issues (e.g. MUC16)**
- **size of the gene (e.g. TTN)**

Why cause missed?



- **family**
  parents consanguineous

image from Gijs Santen

- **index case**
  decreased fetal movement (HP:0001558), severe neonatal hypotonia (HP:0001319), severe feeding difficulties (HP:0011968)

- **exome analysis**
  no causative variants

Imprinting disorders

RESEARCH ARTICLE Open Access

Paraganglioma and pheochromocytoma upon maternal transmission of *SDHD* mutations

Jean-Pierre Bayley[1*†], Rogier A Oldenburg[6†], Jennifer Nuk[8], Attje S Hoekstra[1], Conny A van der Meer[6], Esther Korpershoek[7], Barbara McGillivray[8], Eleonora PM Corssmit[5], Winand NM Dinjens[7], Ronald R de Krijger[7], Peter Devilee[1,3], Jeroen C Jansen[4] and Frederik J Hes[2]

BMC Medical Genetics 2014, 15:111

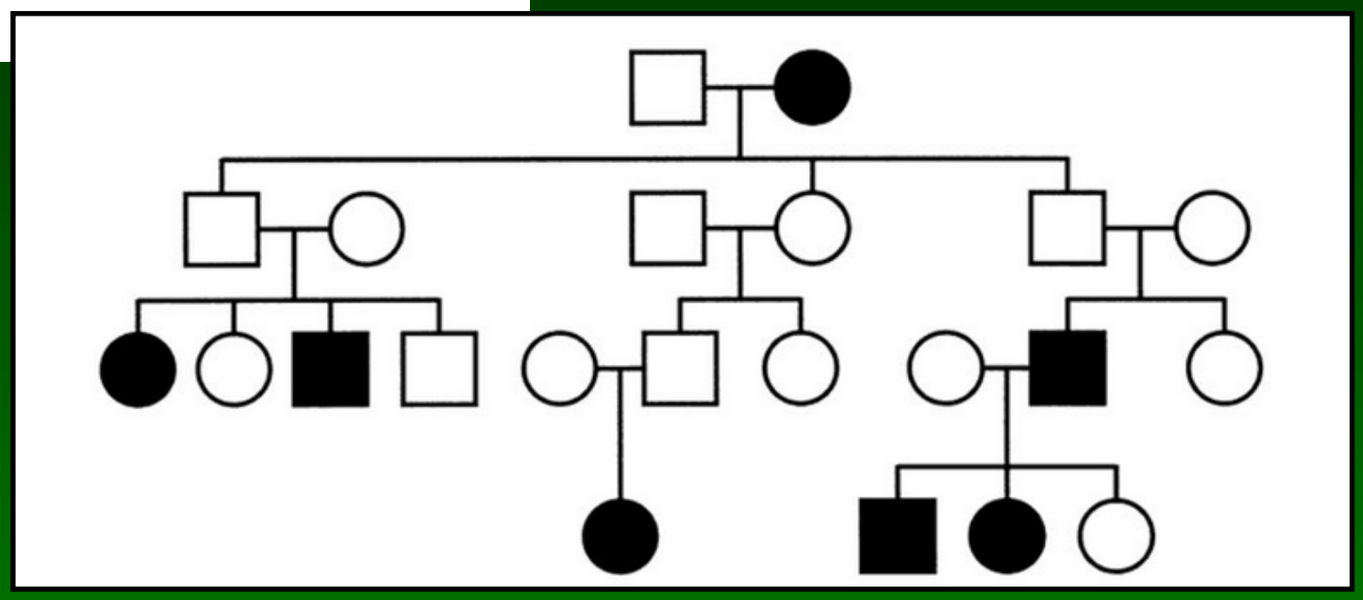


image from Anna Benet-Pages

variant inherited from father: affected
variant inherited from mother: healthy

Reference phenotype

- **phenotype reference genome?**

- **ABO blood group gene**
  genome has no variant called
  ABO blood group type is A, B or O ??

blood group O

so not the "normal" gene

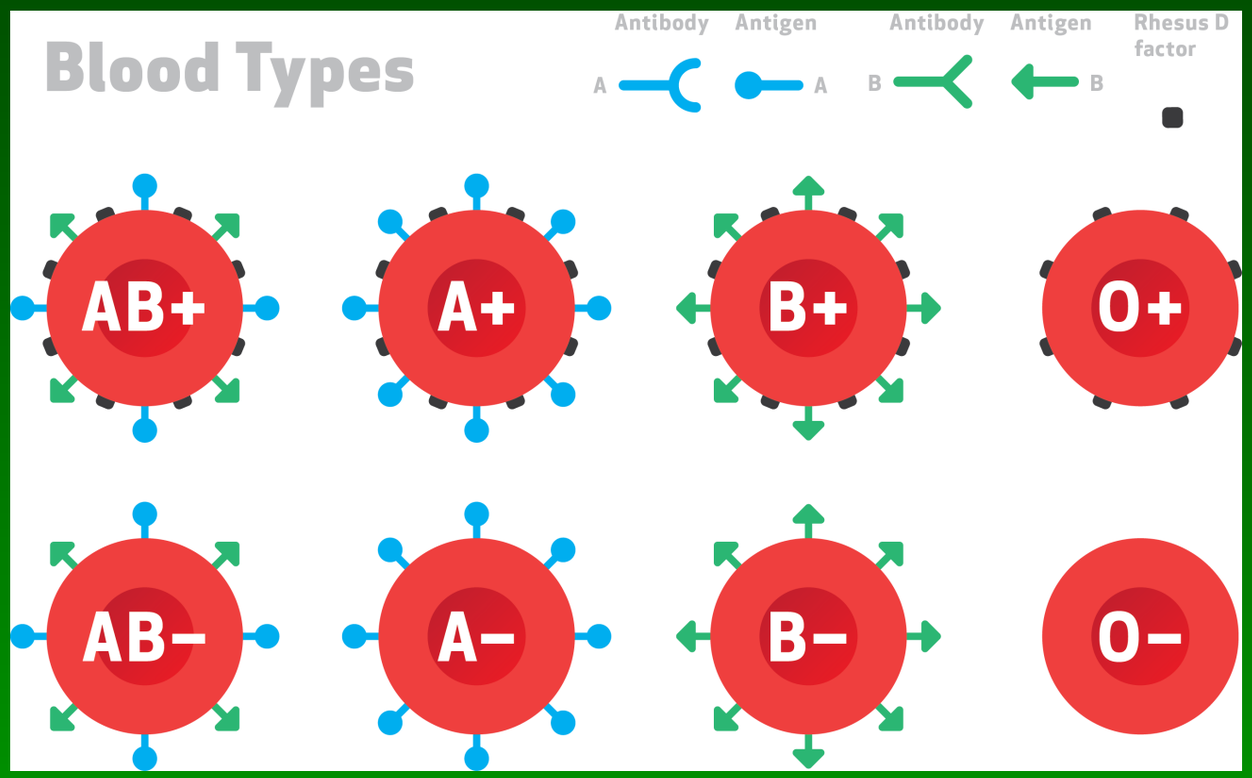


image from www.blood.ca

Reference genome

- not always the major allele
- ref transcript/ref genome may differ
  sometimes with length differences

Reference SNP (refSNP) Cluster Report: rs8176719 ****Clinical Channel****

RefSNP	
Organism:	human (*Homo sapiens*)
Molecule Type:	Genomic
Created/Updated in build:	117/151
Map to Genome Build:	108/Weight 1
Validation Status:	
Citation:	PubMed LitVar NEW
Association:	NHGRI GWAS

Allele	
Variation Class:	DIV: deletion/insertion variation
RefSNP Alleles:	-/G (REV)
Allele Origin:	
Ancestral Allele:	A
Variation Viewer:	VarView
Clinical Significance:	NA
MAF/MinorAlleleCount:	C=0.3764/45143 (ExAC) C=0.3438/1722 (1000 Genomes) C=0.3400/4154 (GO-ESP) C=0.3456/43394 (TOPMED)

HGVS Names

CM000671.2:g.133257521_133257522insC
NC_000009.11:g.136132908_136132909insC
NC_000009.12:g.133257521_133257522insC
NG_006669.1:g.20145_20147insG
NM_020469.2:c.260_262insG
NP_065202.2:p.Val87_Thr88=fs
XP_005276905.1:p.Thr87Aspfs
XP_005276906.1:p.Thr69Aspfs

strange variant descriptions

Alleles -/C | MAF: 0.34 (C) | Highest population MAF: 0.48

Location Chromosome 9: between 133257521 and 133257522 (forward strand) | VCF: 9 133257521 rs8176719 T TC

Evidence status

HGVS names

This variant has 9 HGVS names - Hide

Ensembl HGVS:

- NC_000009.12:g.133257521_133257522insC
- ENST00000453660.4:n.290_291insG
- ENST00000538324.2:c.259-1_259insG
- ENSP00000483018.1:p.Thr87AspfsTer107
- ENST00000611156.4:c.259-1_259insG
- ENSP00000483265.1:p.Thr87AspfsTer107
- ENST00000647353.1:n.54-6370_54-6369insG

dbSNP HGVS:

- NM_020469.2:c.260_262insG
- NP_065202.2:p.Val87_Thr88=fs

Synonyms PharmGKB PA166170086

Original source Variants (including SNPs and indels) imported from dbSNP (release 150) | View in dbSNP

About this variant This variant overlaps 4 transcripts, has 2550 sample genotypes, is associated with 2 phenotypes and is mentioned in 54 citations.

how about g. to c. descriptions ?

Simple sequences

MSH2 : NM_000251:c.942+3A>T

most frequent pathogenic MSH2 variant in HNPCC / Lynch Syndrome

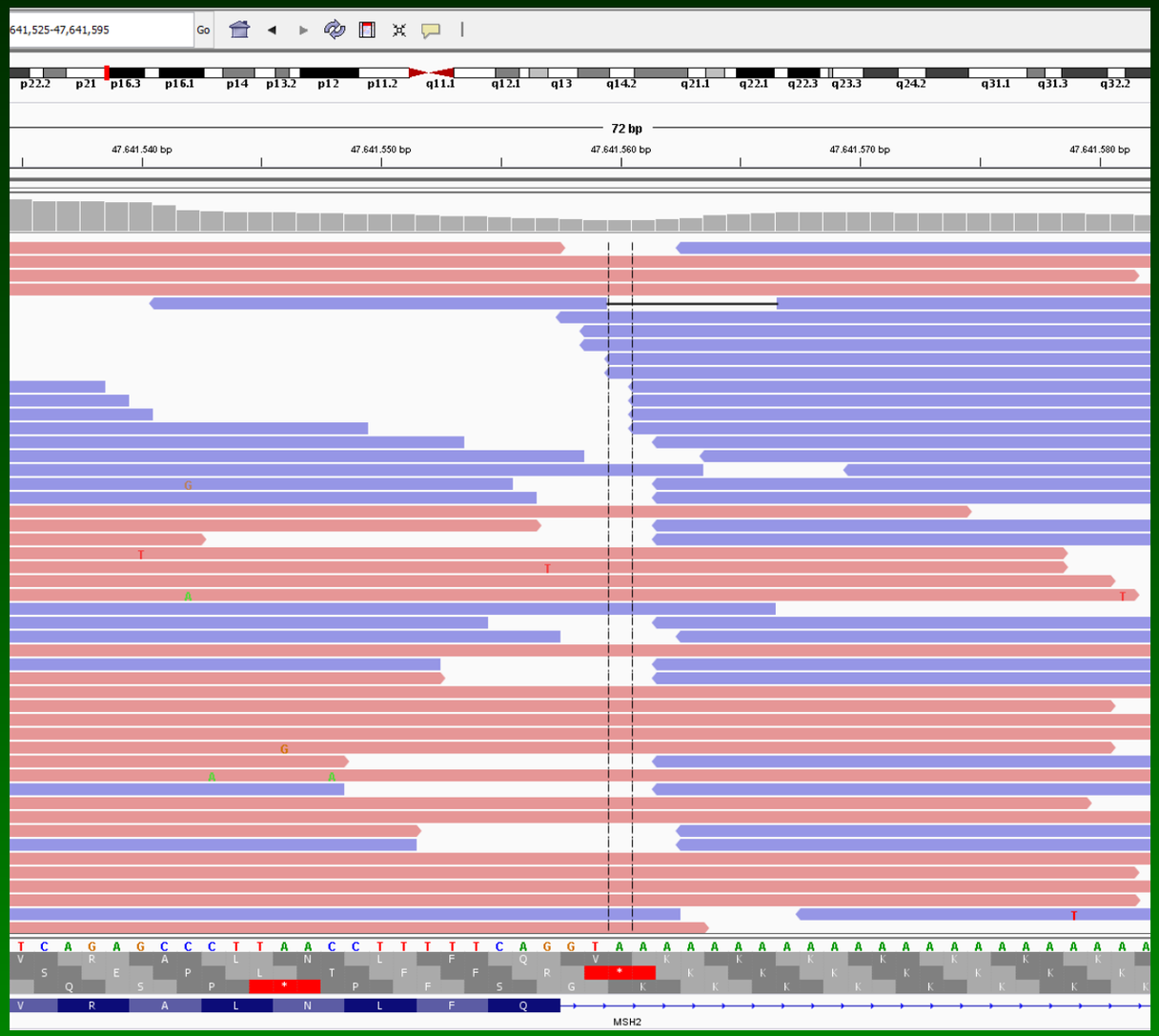
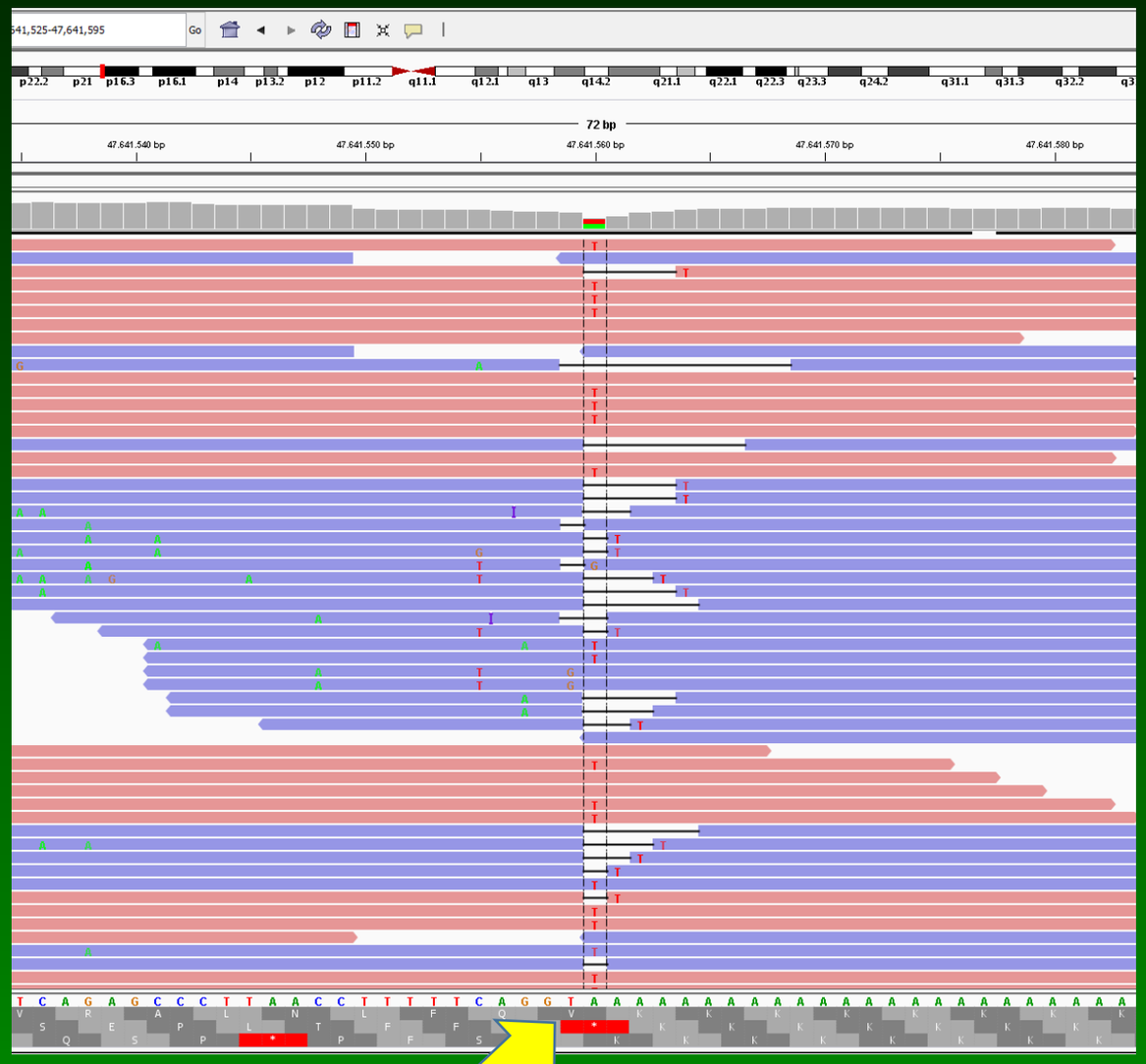


image from Anna Benet-Pages

A



A>T

Somatic variants

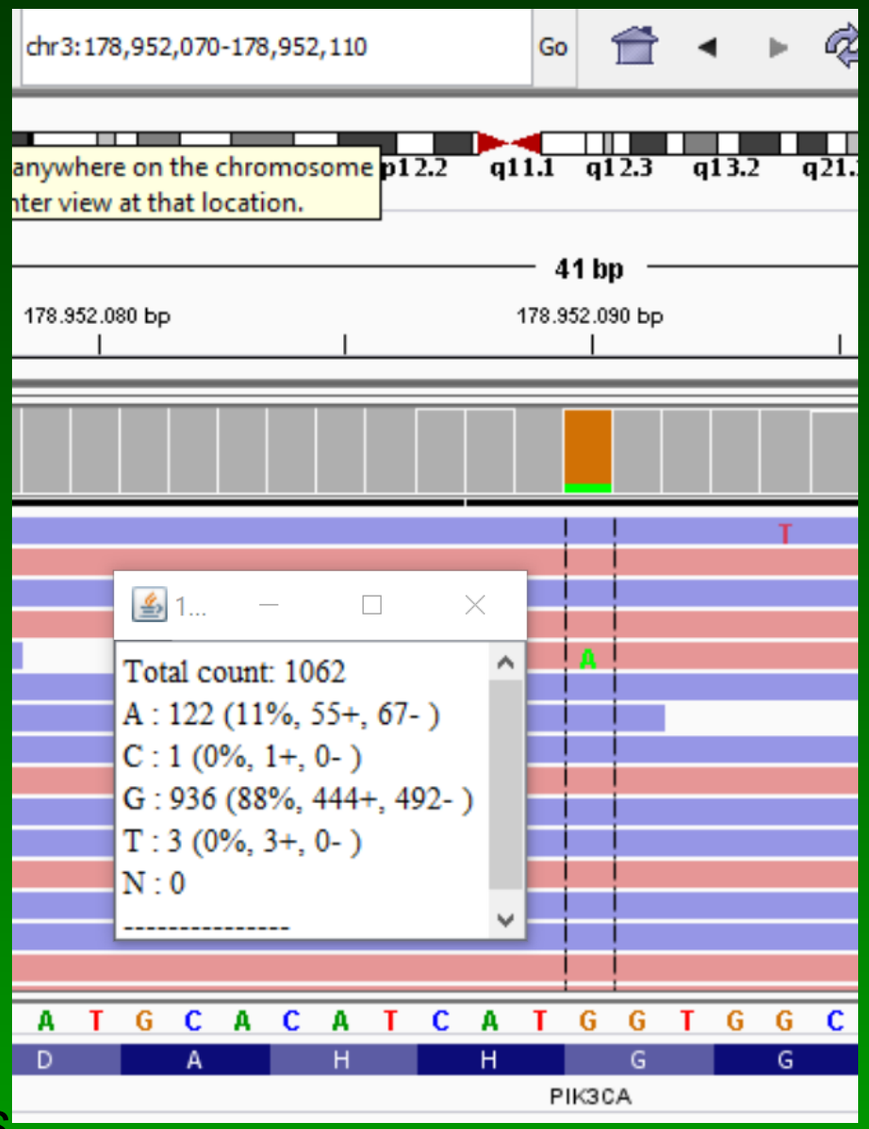
Somatic Mosaic Activating Mutations in *PIK3CA* Cause CLOVES Syndrome

Kyle C. Kurek,[1] Valerie L. Luks,[2] Ugur M. Ayturk,[2,9] Ahmad I. Alomari,[3,6] Steven J. Fishman,[4,6] Samantha A. Spencer,[2,6] John B. Mulliken,[5,6] Margot E. Bowen,[2,9] Guilherme L. Yamamoto,[7] Harry P.W. Kozakewich,[1,6] and Matthew L. Warman[2,6,8,9,*]

only detected with specific LOW-FREQUENCY pipeline

…and when you look for the right variant feature in the right gene/disease

NM_006218.2:c.3145G>A



NM_006218.2:c.2740G>A

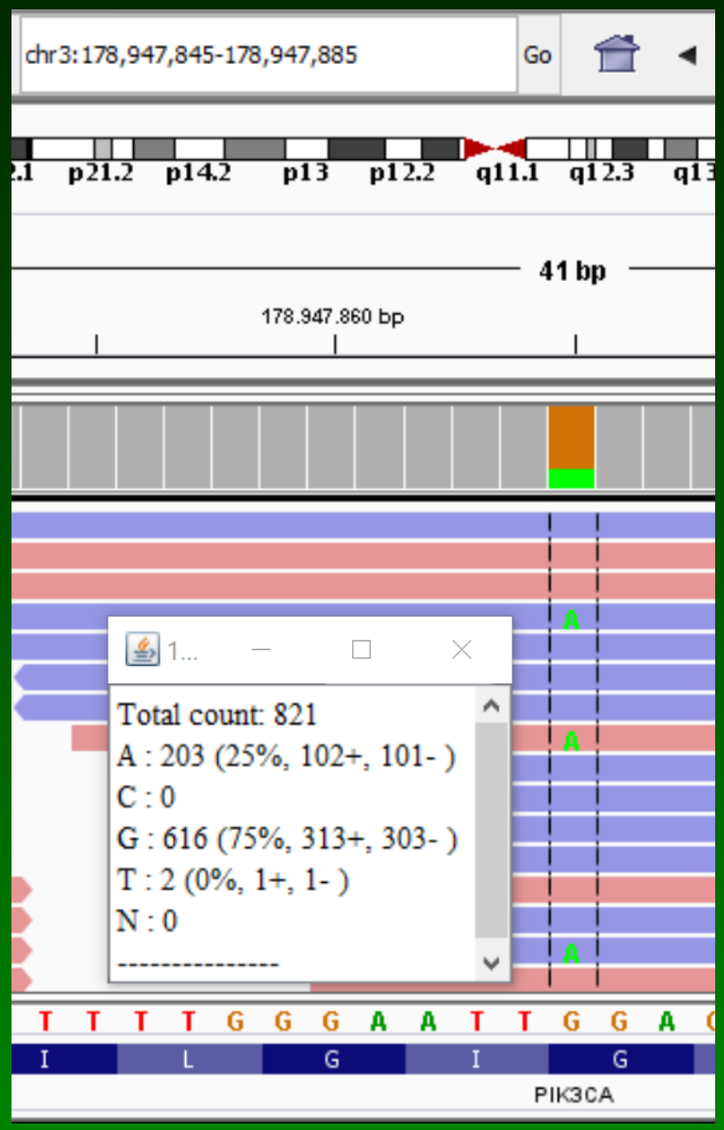


image from Anna Benet-Pages

Variant calling

RASA1

image from Anna Benet-Pages

called: c.1016_1017insTTA p.(Gly340*) and c.1017G>C p.(=)

correct: c.1017delinsTTAC p.(Val39_Gly340insTyr)

be careful when two variant affect one codon

Variant calling

BRCA2

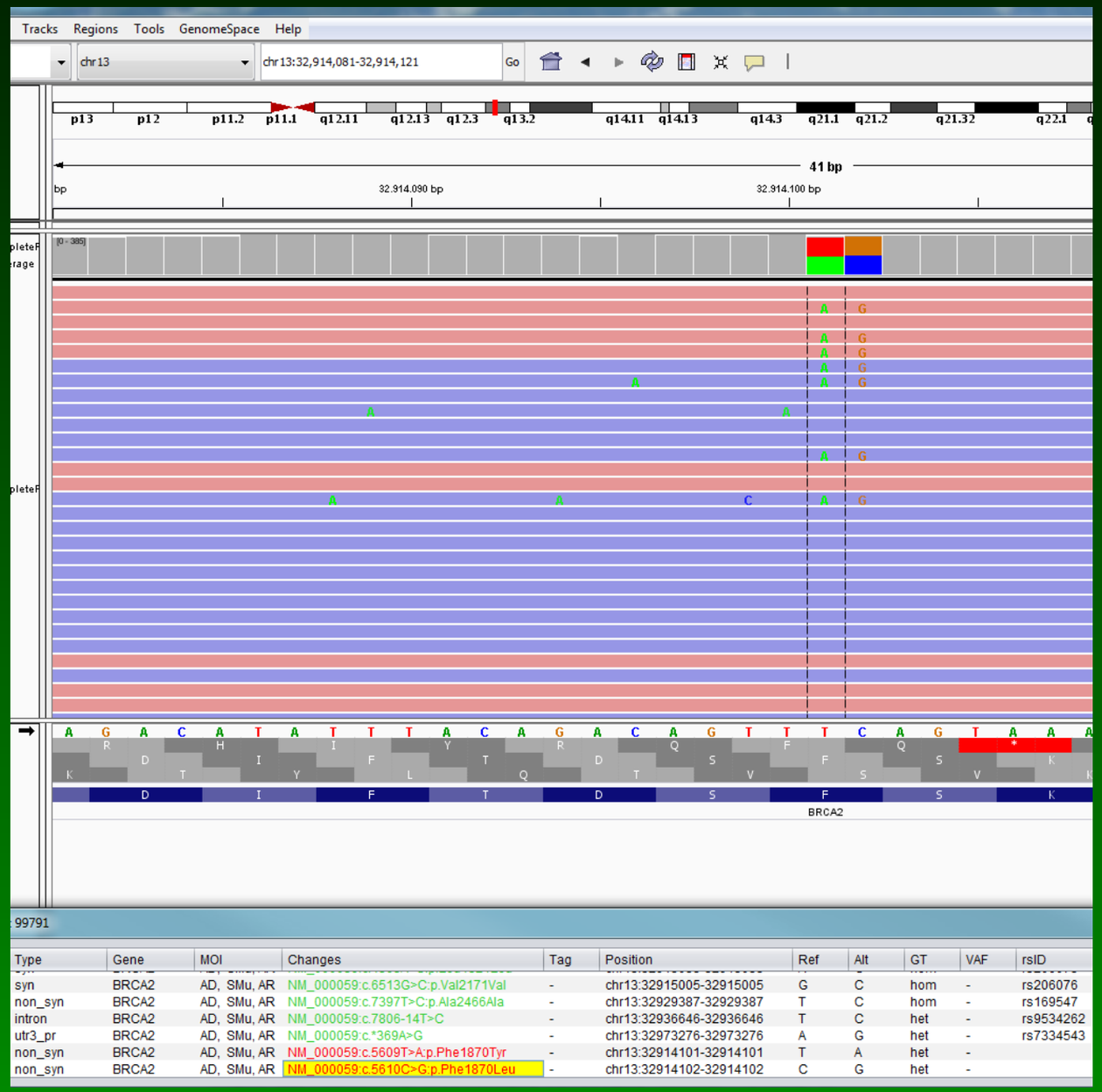


image from Anna Benet-Pages

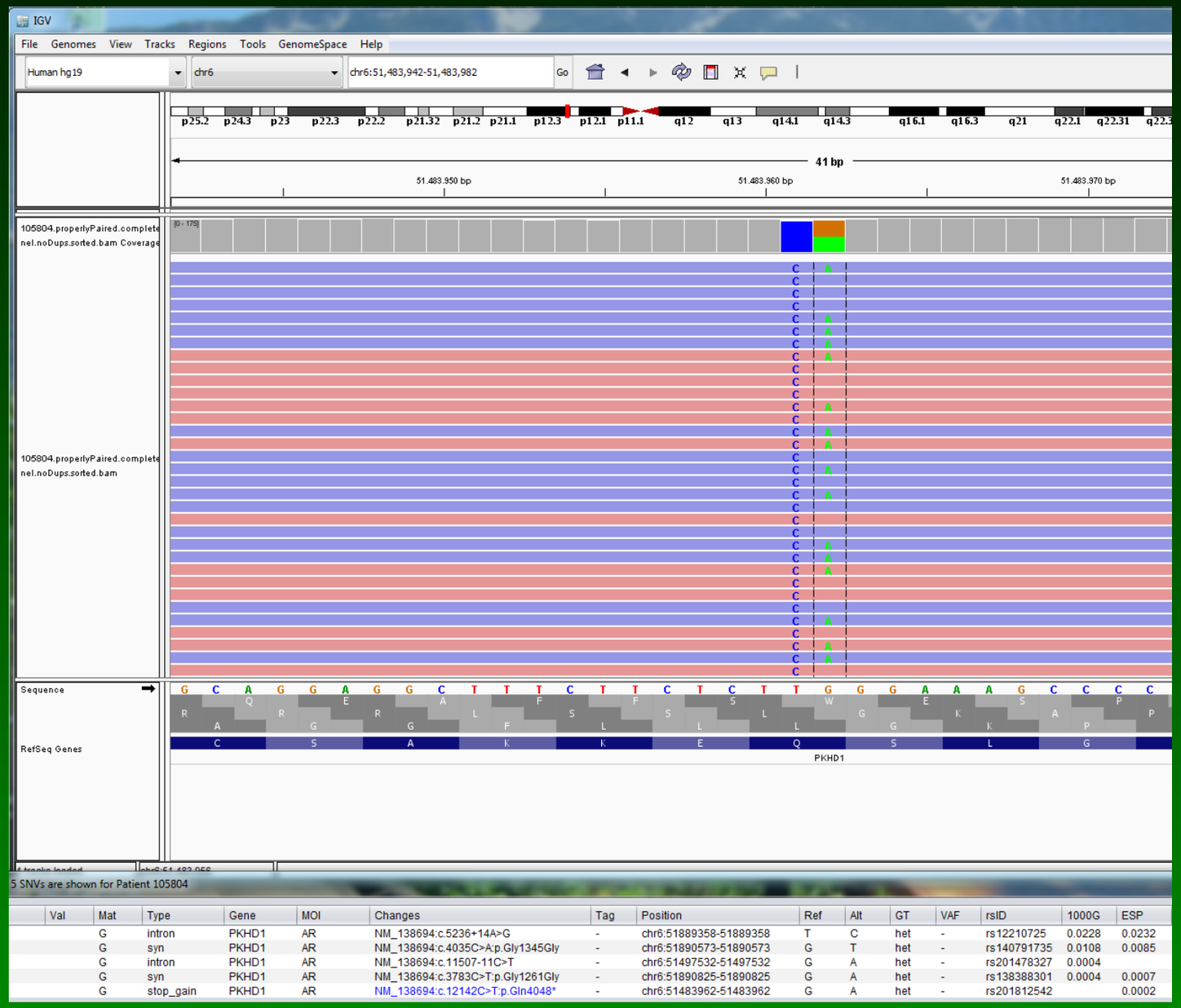
called: c.5609T>A p.(Phe1870Tyr)
and c.5610C>G p.(Phe1870Leu)

correct: c.5609_5610delinsAG
p.(Phe1870*)

How does your pipeline handle such cases ??

Variant calling

PKHD1



Val	Mat	Type	Gene	MOI	Changes	Tag	Position	Ref	Alt	GT	VAF	rsID	1000G	ESP
	G	intron	PKHD1	AR	NM_138694:c.5236+14A>G	-	chr6:51889358-51889358	T	C	het	-	rs12210725	0.0228	0.0232
	G	syn	PKHD1	AR	NM_138694:c.4035C>A:p.Gly1345Gly	-	chr6:51890573-51890573	G	T	het	-	rs140791735	0.0108	0.0085
	G	intron	PKHD1	AR	NM_138694:c.11507-11C>T	-	chr6:51497532-51497532	G	A	het	-	rs201478327	0.0004	
	G	syn	PKHD1	AR	NM_138694:c.3783C>T:p.Gly1261Gly	-	chr6:51890825-51890825	G	A	het	-	rs138388301	0.0004	0.0007
	G	stop_gain	PKHD1	AR	NM_138694:c.12142C>T:p.Gln4048*	-	chr6:51483962-51483962	G	A	het	-	rs201812542		0.0002

image from Anna Benet-Pages

standard pipelines filter out frequent variants (population based)

c.12142C>T p.(Gln4048*)
truncating variant

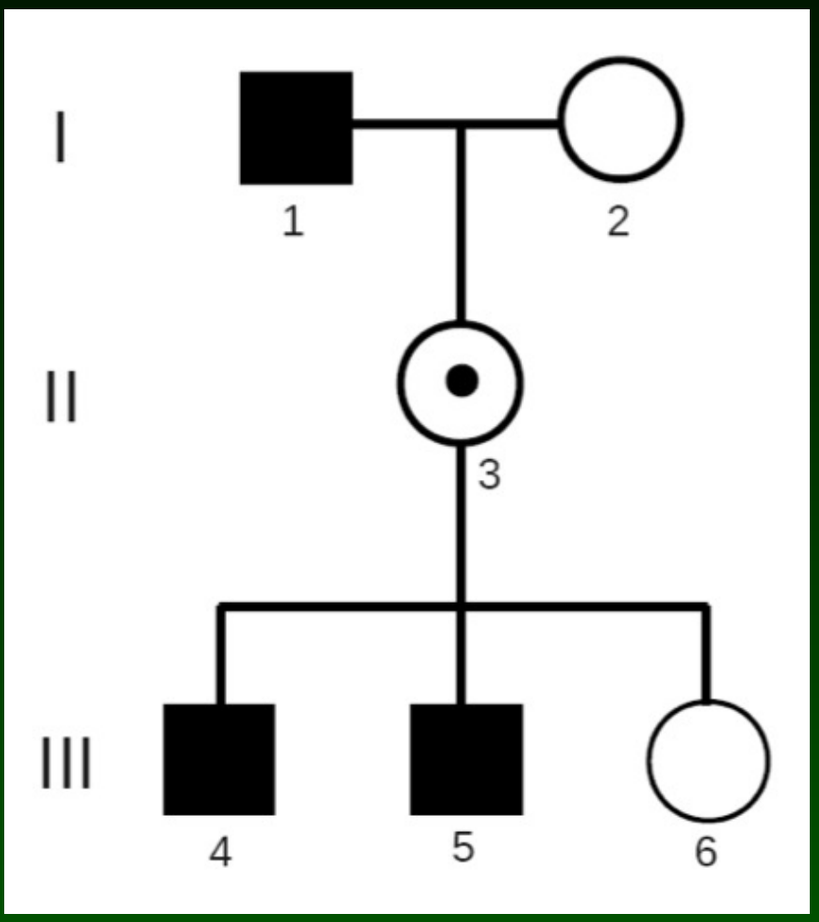
c.12141A>G p.(=)
5% frequency

combination automatically filtered out

Exome focus

- **Aarskog-Scott syndrome**
  FGD1 gene screened
  > no variants

- **whole exome capture**
  no obvious variants
  change thresholds filtering

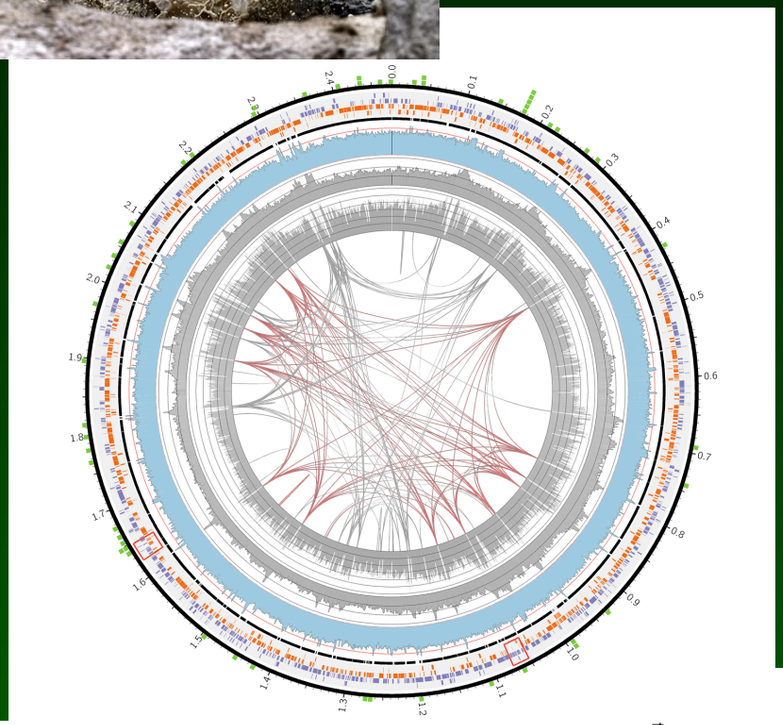


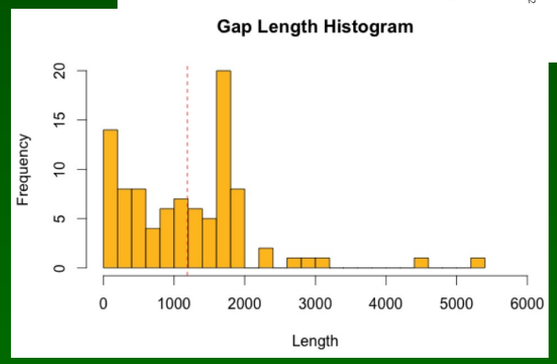
intron -35delA variant

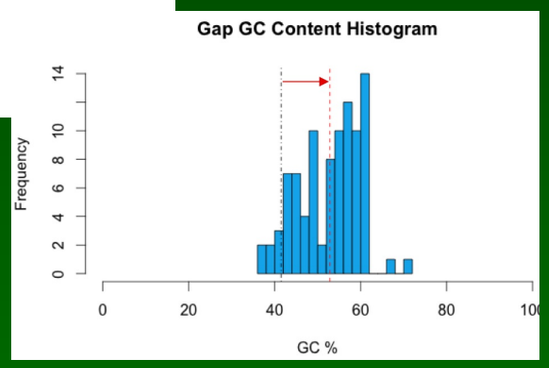
©Yu Sun
Emmelien Aten

Bacterial genome

- sequence across GC-rich regions
- resolved repetitive regions







Anvar *et al. BMC Genomics* 2014, **15**:914
http://www.biomedcentral.com/1471-2164/15/914

BMC Genomics

RESEARCH ARTICLE Open Access

The genomic landscape of the verrucomicrobial methanotroph *Methylacidiphilum fumariolicum* SolV

Seyed Yahya Anvar[1,2*], Jeroen Frank[2], Arjan Pol[3], Arnoud Schmitz[2], Ken Kraaijeveld[2,4], Johan T den Dunnen[1,2,5] and Huub JM Op den Camp[3*]

best result using long reads only

filter reads 8kb or larger

missed 5 kb plasmid

Special cases

KIAA2022 - X-linked
Females (epileptic encephalopathy) / males (intelectual disability)

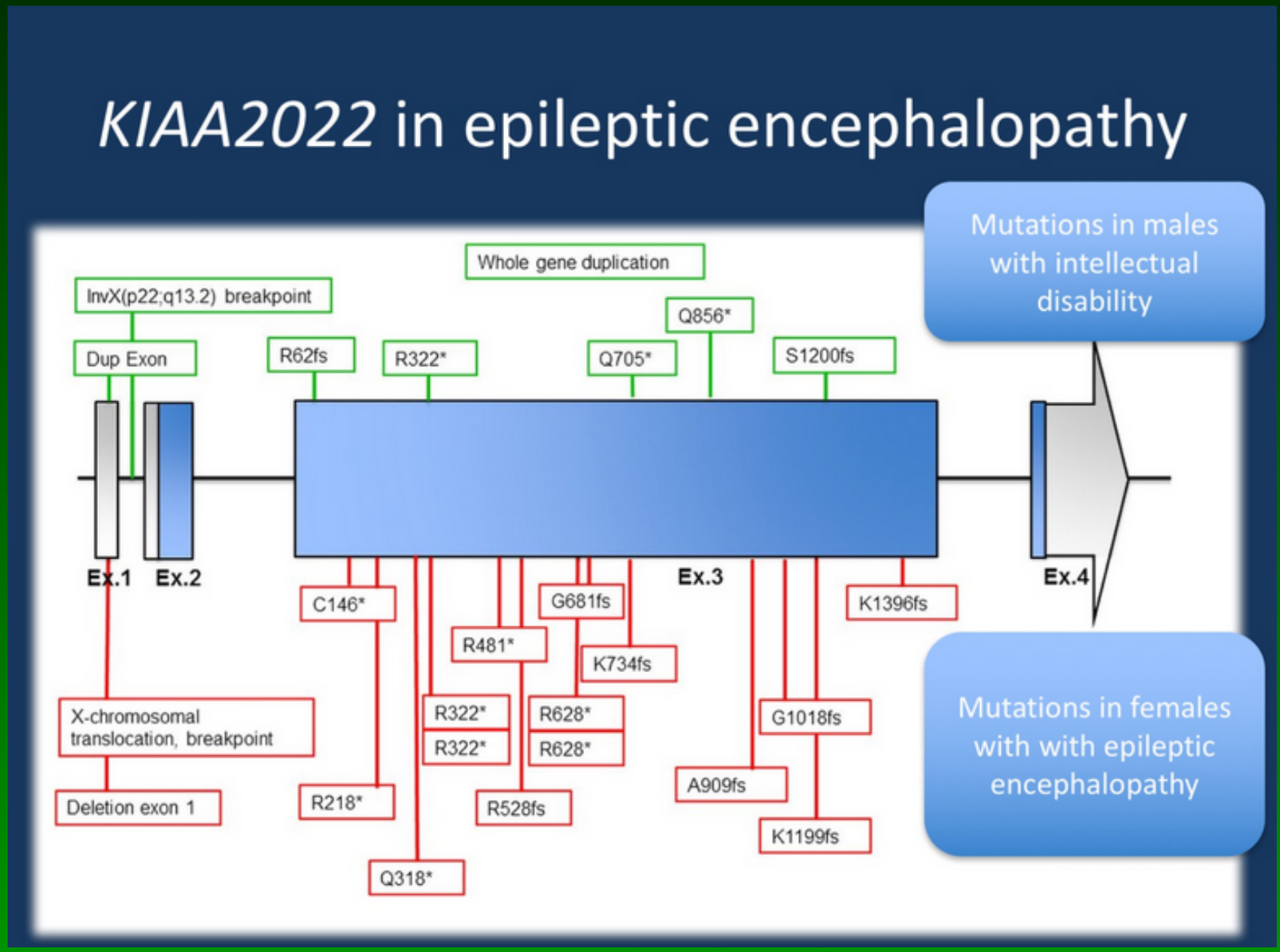
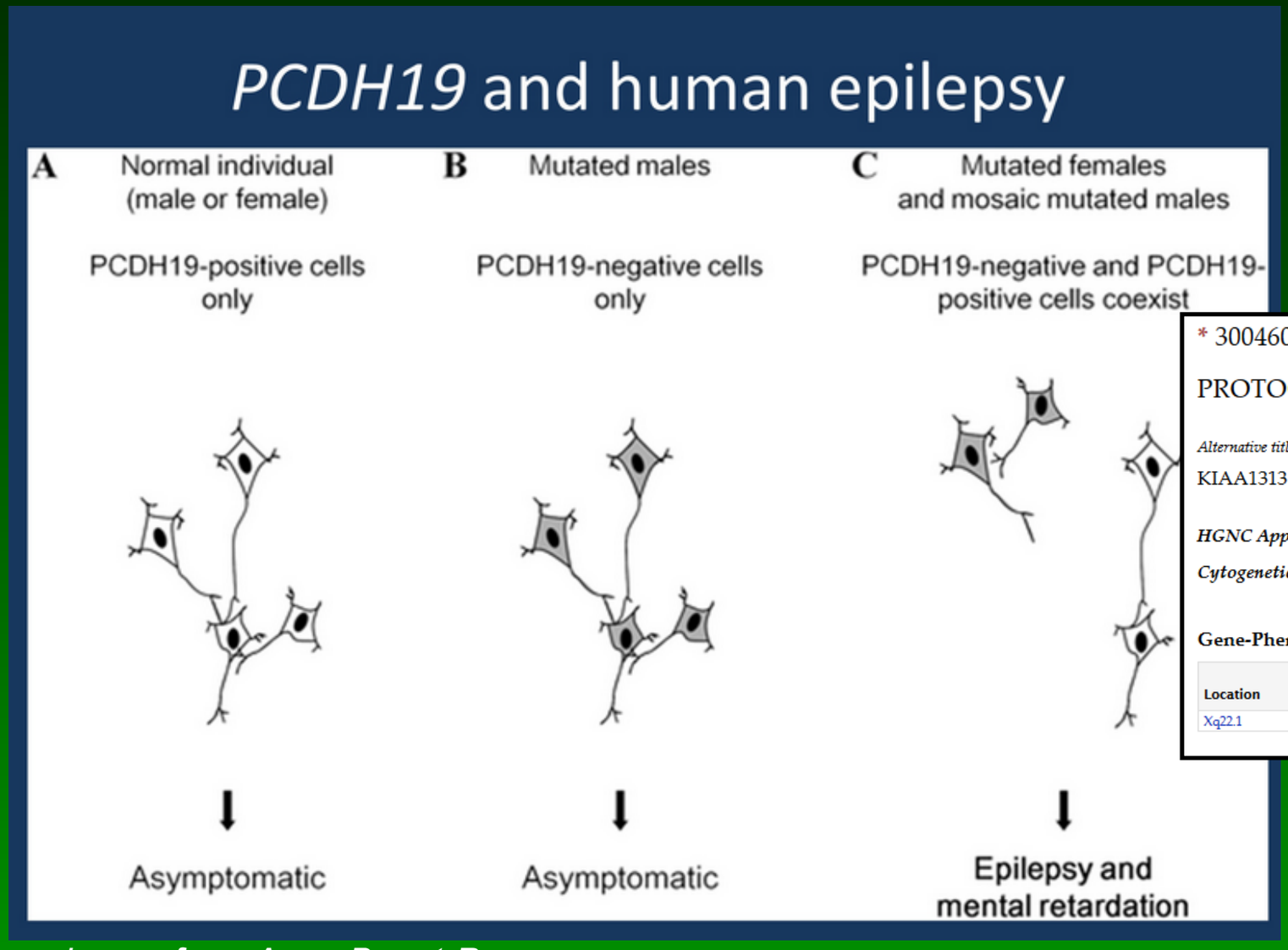
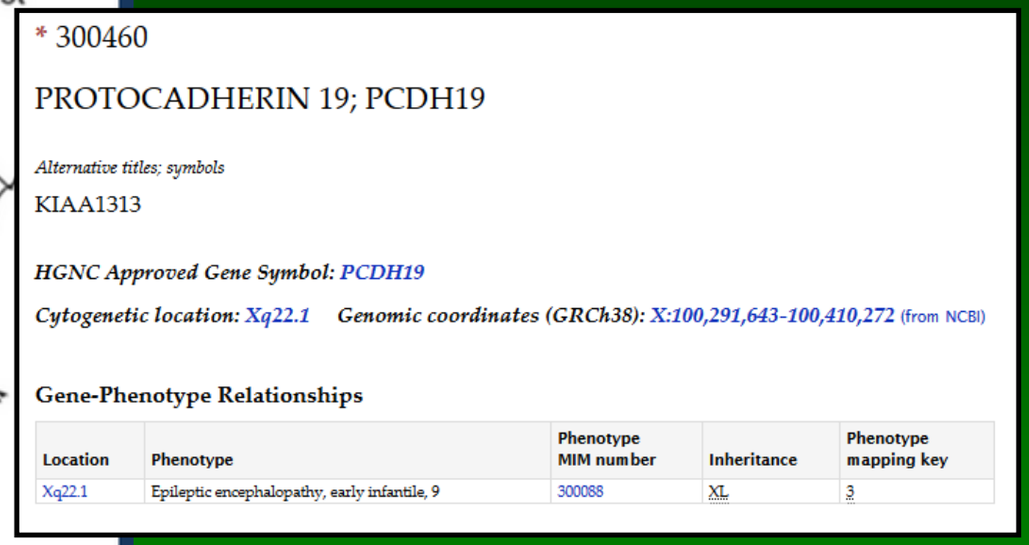


image from Anna Benet-Pages

Special cases

PCDH19: X-linked
… but ONLY heterozygous females and mosaic males are affected





* 300460

PROTOCADHERIN 19; PCDH19

Alternative titles; symbols

KIAA1313

HGNC Approved Gene Symbol: ***PCDH19***

Cytogenetic location: ***Xq22.1*** ***Genomic coordinates (GRCh38):*** ***X:100,291,643-100,410,272*** (from NCBI)

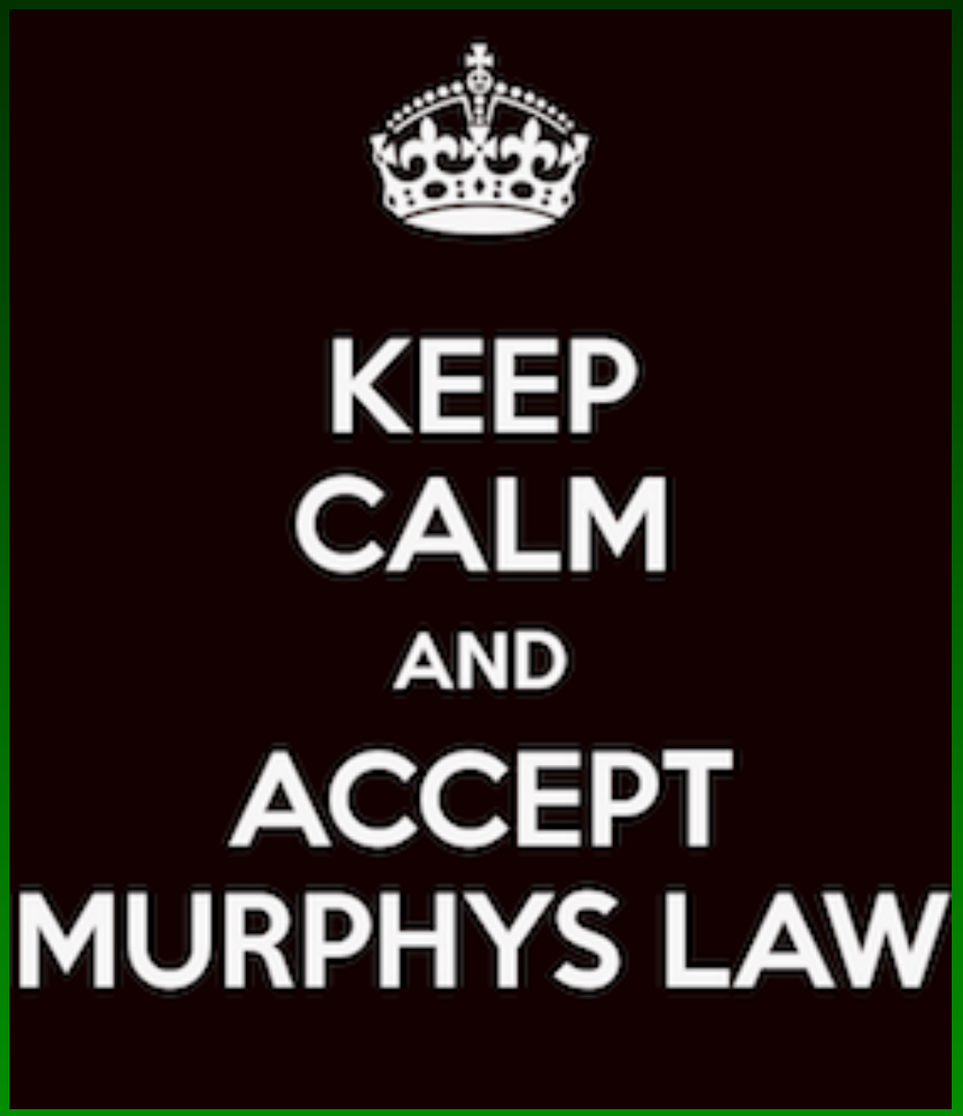
Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
Xq22.1	Epileptic encephalopathy, early infantile, 9	300088	XL	3

image from Anna Benet-Pages

Murphy's law

…also applies in NGS



Acknowledgement

Presentation prepared by:

Johan den Dunnen

Human Genetics & Clinical Genetics
Leiden University Medical Center
Leiden, Nederland

date: April 2019

with contributions from Anna Benet-Pages, Christian Gilissen and Gijs Santen