Reference genomes and Gene models
Impact on variant interpretation

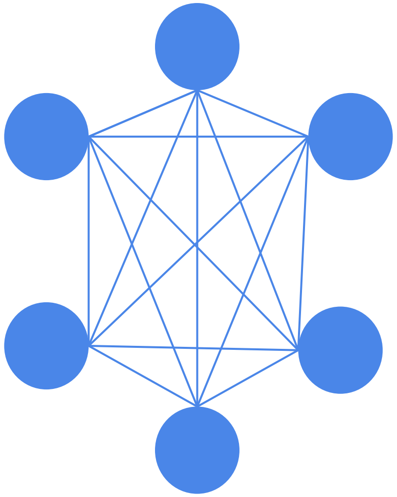
Equipe Bioinformatique
CHU de Lille

- **Part 1 - Reference Genome**

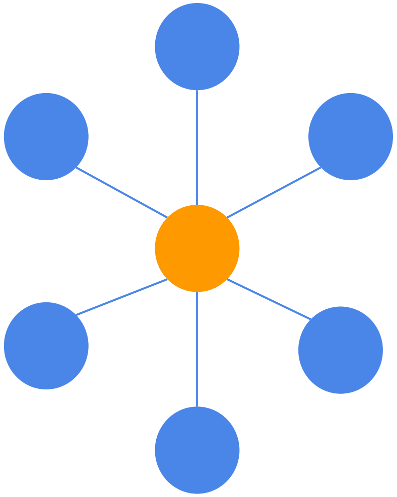
1. The need of a reference
2. Do you speak reference genomes?
3. Human reference genome build : a 3 billion pieces puzzle
4. One given version, but so many flavors
5. Impact on data analysis

The need of a reference

Comparing genome sequences



Complexity +++



Use of a common standard

The need of a reference

Reference genome sequence should be:

- Representative of the human population diversity
- Each segment = most commonly observed across available individual genomes.
- No one's genome, and hopefully everyone's.

The quick brown f**a**x jumped over the lazy doge.

The quick _ fox jump**s** over the lazy doge.

The quick brown fox jump**s** over the lazy **brown** dog.

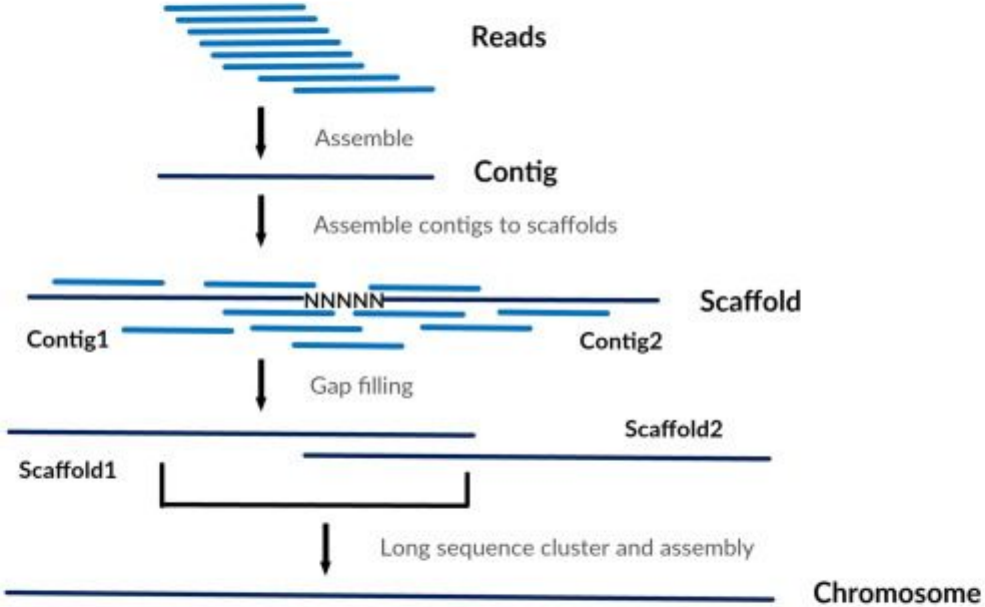
↓

REFERENCE : The quick brown fox jumped over the lazy doge.

Modified from https://software.broadinstitute.org/gatk/documentation/article?id=11011

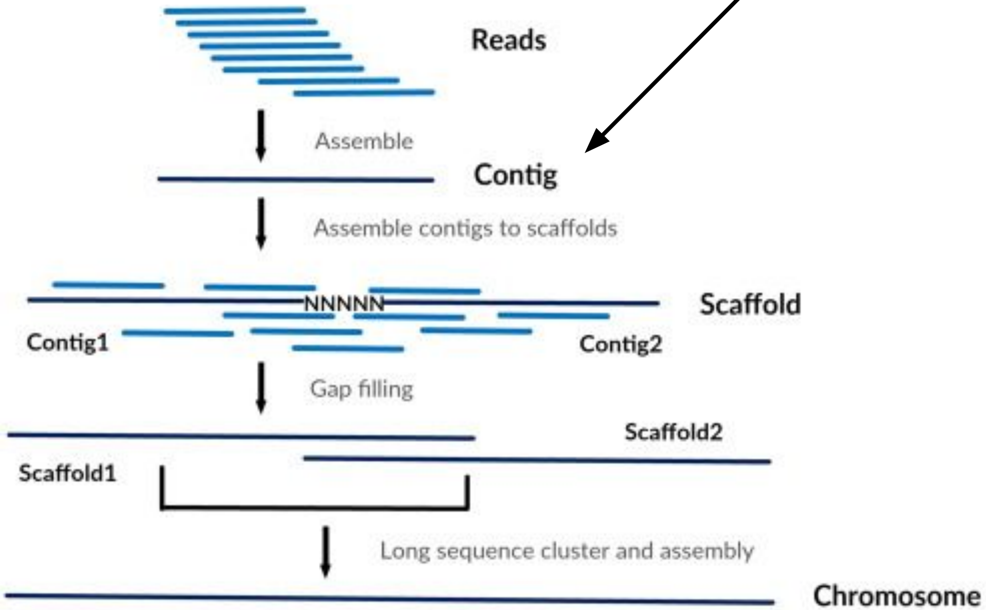
Do you speak ` reference genomes ` ?

Reference genome construction



Do you speak ` reference genomes ` ?

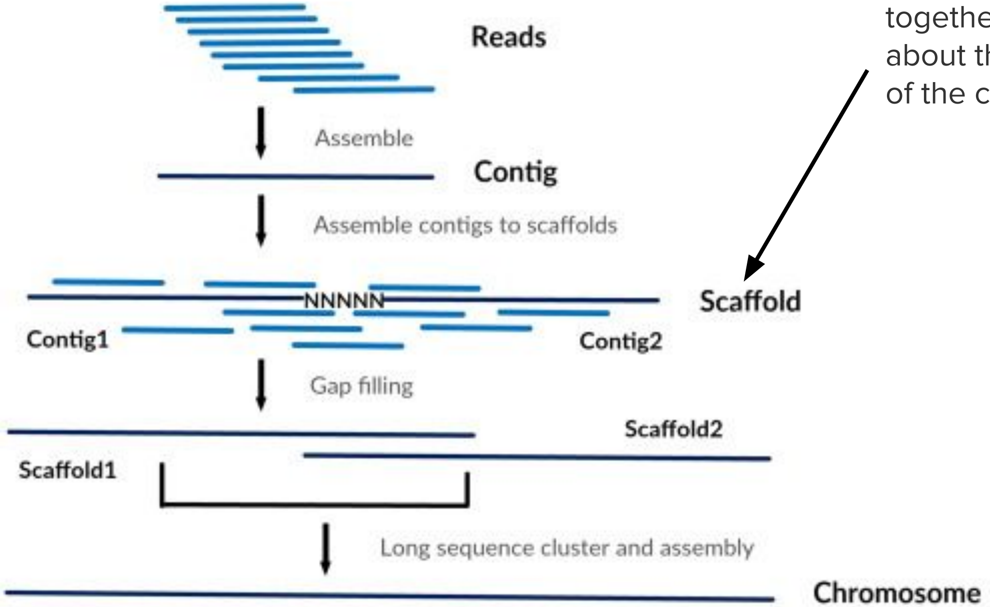
Reference genome construction



Contigs : continuous stretches of sequence containing only A, C, G, or T bases without "physical" gaps. The order of bases is known to a high confidence level.

Do you speak ` reference genomes ` ?

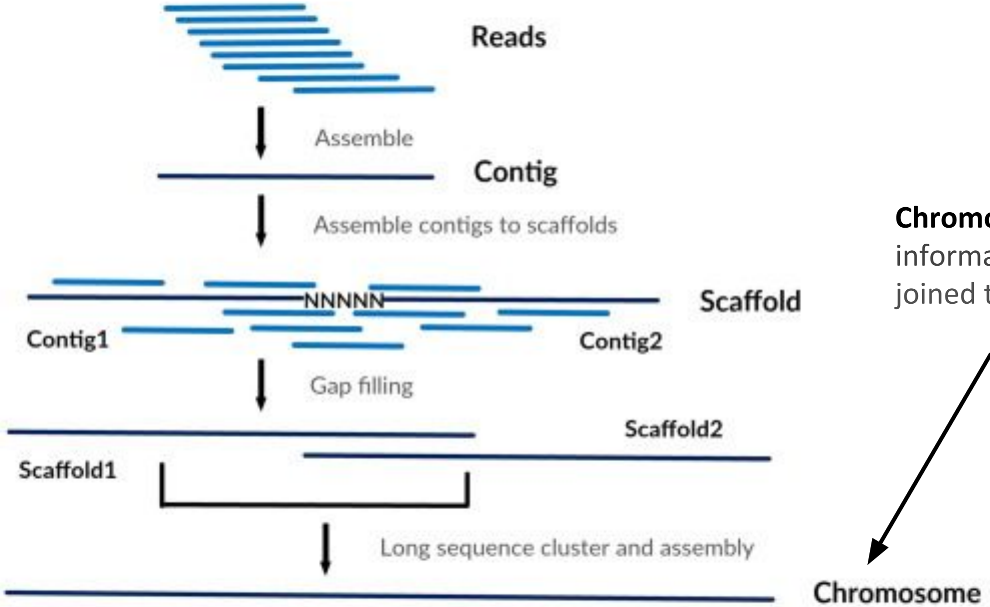
Reference genome construction



Scaffolds : created by chaining contigs together using additional information about the relative position and orientation of the contigs in the genome.

Do you speak ` reference genomes ` ?

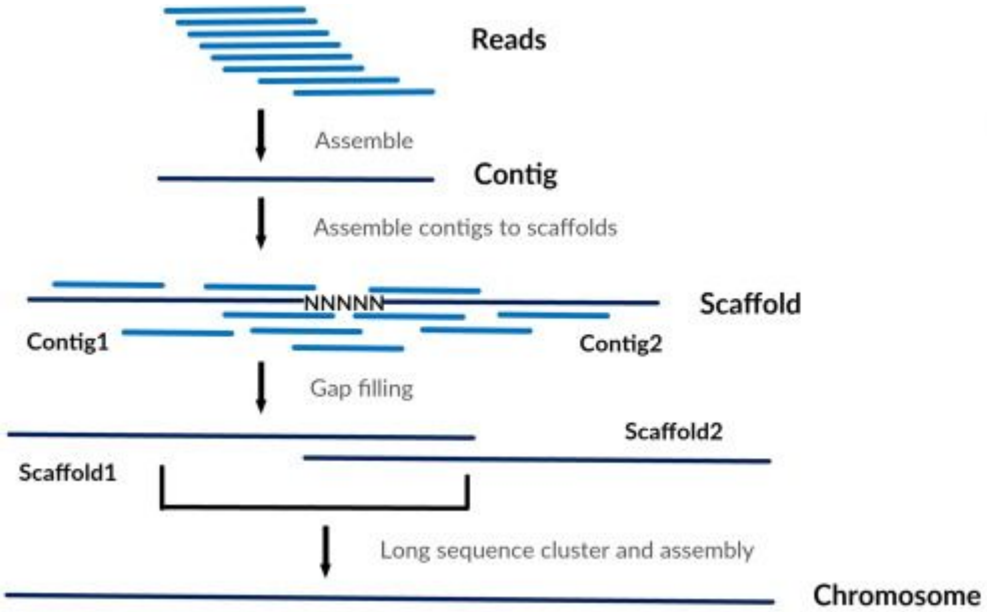
Reference genome construction



Chromosome : when sufficient mapping information, multiple scaffolds can be joined together to form a chromosome

Do you speak ` reference genomes ` ?

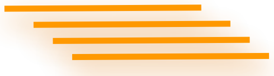
Reference genome construction



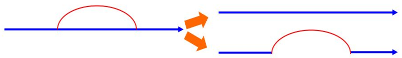
After chromosome assembly, some scaffolds remain.

These are specific cases:

Unplaced scaffolds : not associated with any chromosome.



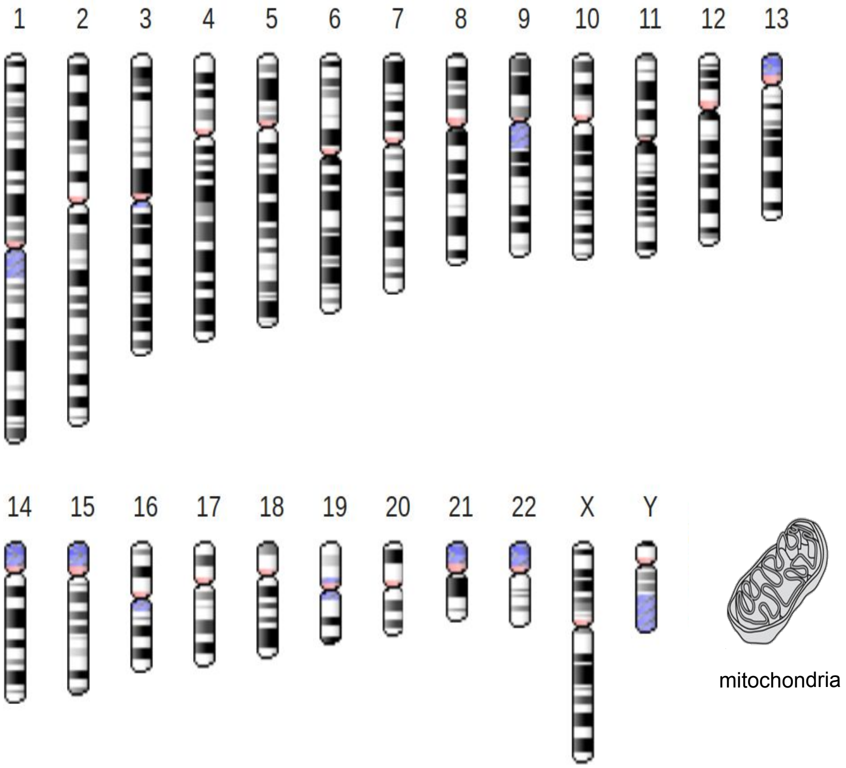
Unlocalised scaffolds : associated with a specific chromosome but cannot be ordered or oriented.



Alternate Loci : representation of diverging haplotypes in regions that are too complex for a single representation

Do you speak ` reference genomes ` ?

Top-level Assembly =

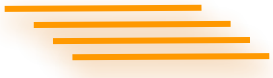


Assembled chromosomes

+

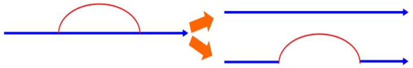
Unplaced scaffolds

+



Unlocalised scaffolds

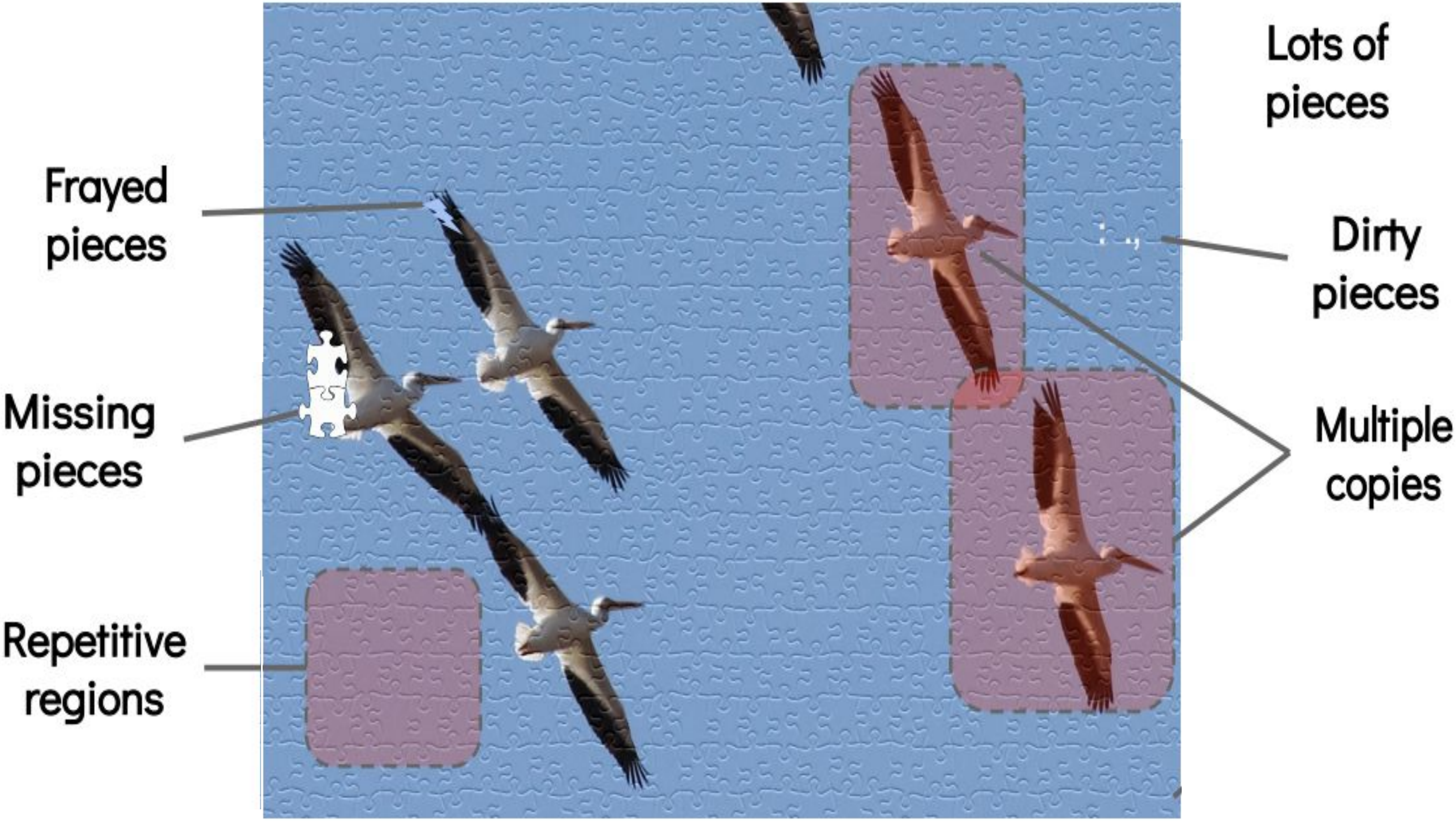
+



Alternate Loci

Human reference genome build : a 3 billion pieces puzzle

Human reference genome build : a 3 billion pieces puzzle

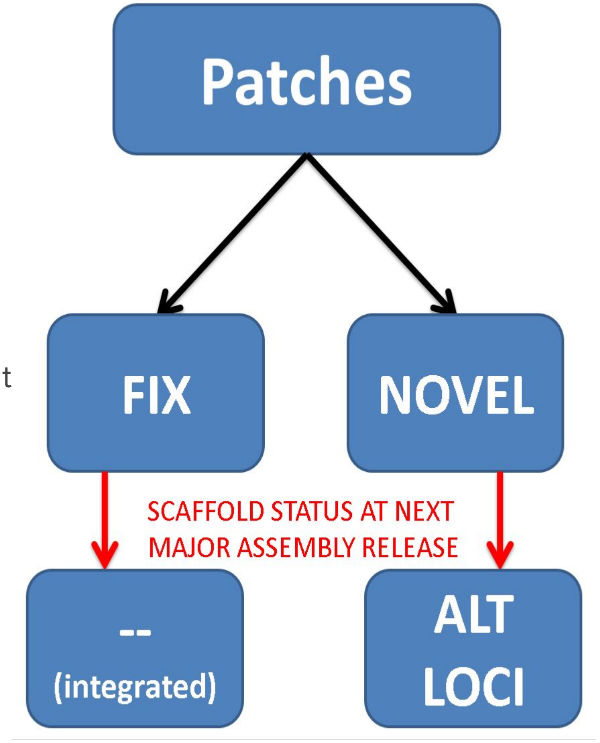


Human reference genome build : a 3 billion pieces puzzle

An assembly is never perfect, but in constant progress…

Patches = assembly updates, not disrupting the chromosome coordinates

FIX = error corrections (base changes, component replacements, ..)



Patches

FIX

NOVEL

SCAFFOLD STATUS AT NEXT MAJOR ASSEMBLY RELEASE

--
(integrated)

ALT LOCI

NOVEL = addition of new alternate loci to the assembly

Human reference genome build : a 3 billion pieces puzzle

An assembly is never perfect, but in constant progress…

Patches release # Minor release

GRCh37.p1 => GRCh37.p2 => … GRCh37.p12 => GRCh37.p13

Genome assembly release # Major release

GRCh37 => GRCh38

One given version, but so many flavors

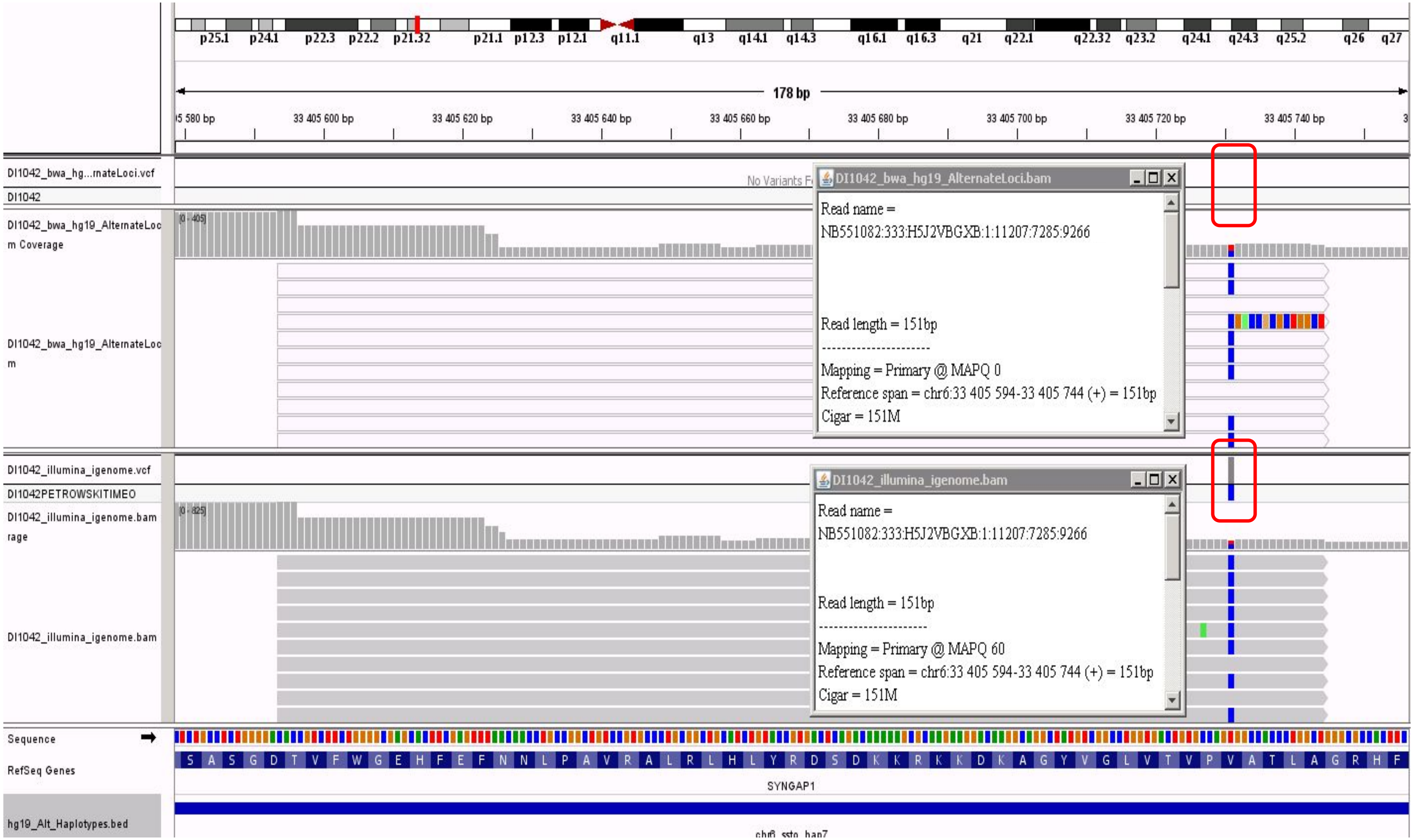
Flavor	Source	Name	Unplaced contigs	Unlocalized contigs	Alternate loci	mitochondria	Epstein-Barr virus	decoy sequences	Remarks
GRCH		GRCh37	No canonical name	No canonical name	No canonical name	Maintained by Mitomap, distributed for convenience	✗	✗	
UCSC	GRCh37	hg19	chrUn_gl000212	chr1_gl000191_random	chr6_apd_hap1	NC_001807 (from build 36)	✗	✗	Chromosome names start by “chr” PAR regions on chrY are hard masked
Ensembl	GRCh37.p13	Ensembl API release 75 Homo_sapiens.GRCh37.75.dna.primary_assembly.fasta.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	Chromosome named "1" to "22", "X", "Y" and "MT"
1000 genomes project phase I & III	GRCh37.p2	hs37 g1k_v37 b37 human_g1k_v37.fasta.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	"1" to "22", "X", "Y" and "MT"

One given version, but so many flavors

Flavor	Source	Name	Unplaced contigs	Unlocalized contigs	Alternate loci	mitochondria	Epstein-Barr virus	decoy sequences	Remarks
1000 genomes project phase II	GRCh37. p4	hs37d5 b37+decoy +herpes hs37d5.fa.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	NC_00 7605	hs37d5 ss	pseudo-autosomal regions are hard-marked on Y chromosome
Illumina MiSeq Reporter + BSO	hg19	hg19	✗	✗	✗	NC_001807 (from build 36)	✗	✗	hg19 without unplaced/unlocalized contigs nor alternate loci
Ion Torrent	hg19	hg19	✗	✗	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	hg19 without unplaced/unlocalized contigs nor alternate loci
GATK Bundle	GRCh37. p2	b37 + decoy	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	hs37d5 ss	"1" to "22", "X", "Y" and "MT"

Impact on data analysis

ALT contigs : Mapping quality zero for reads mapped in the flanking sequences.
Sensitivity of variant calling ↘↘
ALT-aware mapper


p25.1 p24.1 p22.3 p22.2 p21.32 p21.1 p12.3 p12.1 q11.1 q13 q14.1 q14.3 q16.1 q16.3 q21 q22.1 q22.32 q23.2 q24.1 q24.3 q25.2 q26 q27
178 bp
5 580 bp
33 405 600 bp
33 405 620 bp
33 405 640 bp
33 405 660 bp
33 405 680 bp
33 405 700 bp
33 405 720 bp
33 405 740 bp
DI1042_bwa_hg...rnateLoci.vcf
DI1042
No Variants F
DI1042_bwa_hg19_AlternateLoc m Coverage
[0 - 405]
DI1042_bwa_hg19_AlternateLoc m
DI1042_illumina_igenome.vcf
DI1042PETROWSKITIMEO
DI1042_illumina_igenome.bam rage
[0 - 825]
DI1042_illumina_igenome.bam
Sequence
RefSeq Genes
S A S G D T V F W G E H F E F N N L P A V R A L R L H L Y R D S D K K R K K D K A G Y V G L V T V P V A T L A G R H F
SYNGAP1
hg19_Alt_Haplotypes.bed
chr6_ssto_hap7

DI1042_bwa_hg19_AlternateLoci.bam

Read name =
NB551082:333:H5J2VBGXB:1:11207:7285:9266

Read length = 151bp

Mapping = Primary @ MAPQ 0
Reference span = chr6:33 405 594-33 405 744 (+) = 151bp
Cigar = 151M

DI1042_illumina_igenome.bam

Read name =
NB551082:333:H5J2VBGXB:1:11207:7285:9266

Read length = 151bp

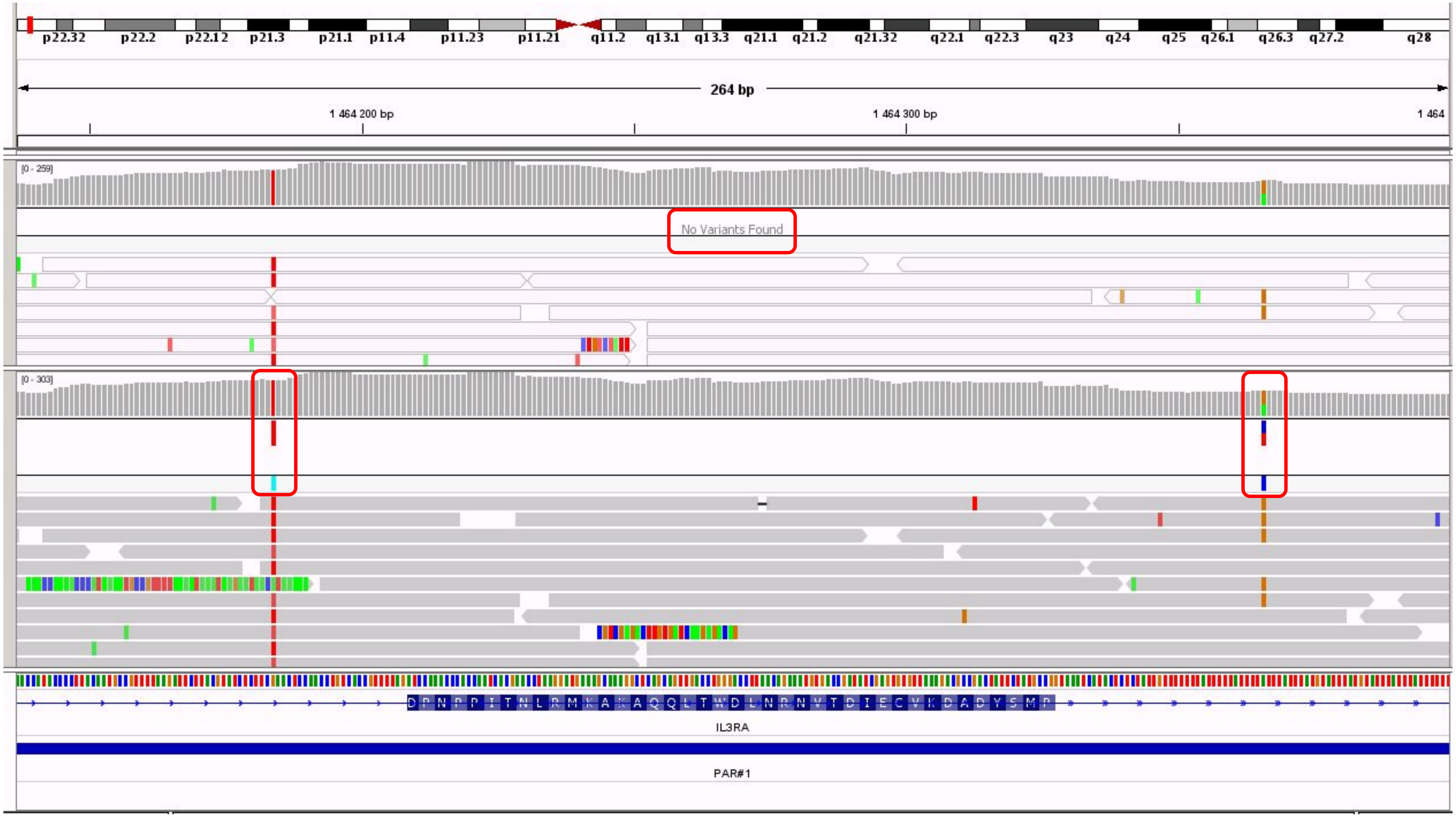
Mapping = Primary @ MAPQ 60
Reference span = chr6:33 405 594-33 405 744 (+) = 151bp
Cigar = 151M

Impact on data analysis

Multi-placed sequences : Pseudo-autosomal regions (PARs).

If placed on both chrX and chrY, standard pipeline not be able to call any variants in PARs.

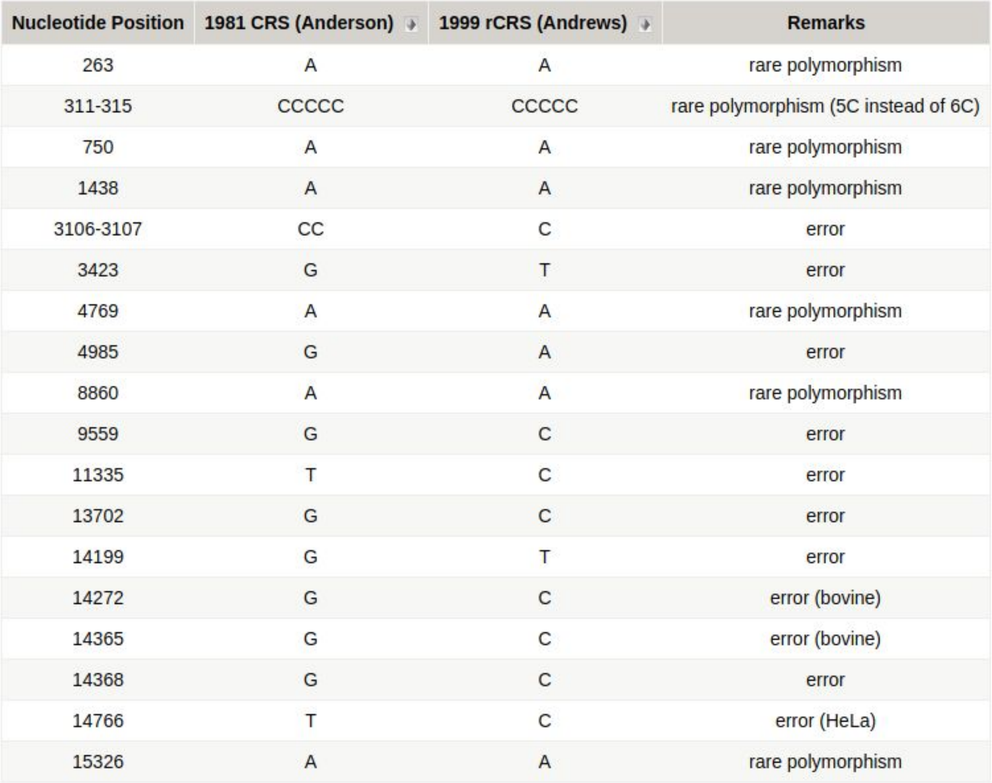
Solution = hard mask PARs on chrY.



Impact on data analysis

Not using the rCRS mitochondrial sequence (NC_012920.1)

NC_001807 : wrong sequence length + 2 bp insertion



Nucleotide Position	1981 CRS (Anderson)	1999 rCRS (Andrews)	Remarks
263	A	A	rare polymorphism
311-315	CCCCC	CCCCC	rare polymorphism (5C instead of 6C)
750	A	A	rare polymorphism
1438	A	A	rare polymorphism
3106-3107	CC	C	error
3423	G	T	error
4769	A	A	rare polymorphism
4985	G	A	error
8860	A	A	rare polymorphism
9559	G	C	error
11335	T	C	error
13702	G	C	error
14199	G	T	error
14272	G	C	error (bovine)
14365	G	C	error (bovine)
14368	G	C	error
14766	T	C	error (HeLa)
15326	A	A	rare polymorphism

7 nucleotides as rare polymorphisms

+

11 corrected errors

Impact on data analysis

Decoy sequences =

- sequences derived from HuRef, Human Bac and Fosmid clones and NA12878
- known true human genome sequences which are not in the reference genome sequence.

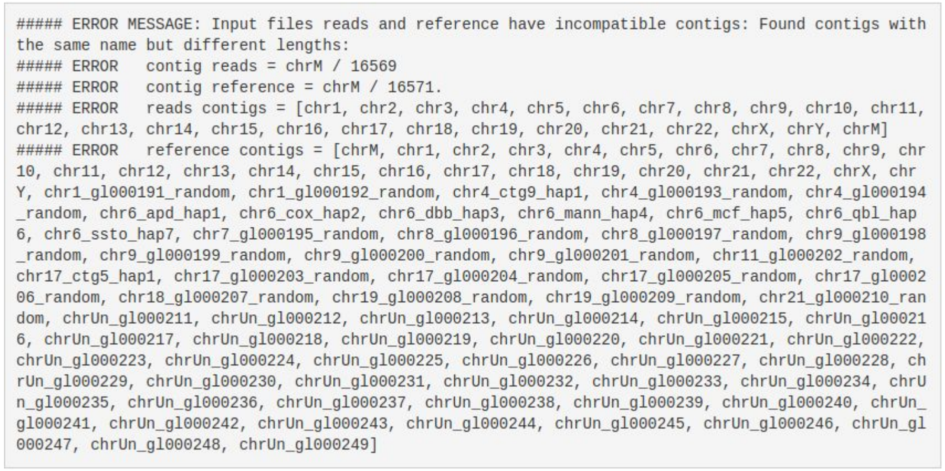
- many reads will quickly find a very confident alignment in the decoy
- If absent, reads would otherwise map with low quality on the reference genome sequence

Mapping process speed ↗️↗️

False positive calls ↘️

Impact on data analysis

Comparing/Combining your own data with external files from collaborators



```
##### ERROR MESSAGE: Input files reads and reference have incompatible contigs: Found contigs with
the same name but different lengths:
##### ERROR   contig reads = chrM / 16569
##### ERROR   contig reference = chrM / 16571.
##### ERROR   reads contigs = [chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr10, chr11,
chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrX, chrY, chrM]
##### ERROR   reference contigs = [chrM, chr1, chr2, chr3, chr4, chr5, chr6, chr7, chr8, chr9, chr
10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr19, chr20, chr21, chr22, chrX, chr
Y, chr1_gl000191_random, chr1_gl000192_random, chr4_ctg9_hap1, chr4_gl000193_random, chr4_gl000194
_random, chr6_apd_hap1, chr6_cox_hap2, chr6_dbb_hap3, chr6_mann_hap4, chr6_mcf_hap5, chr6_qbl_hap
6, chr6_ssto_hap7, chr7_gl000195_random, chr8_gl000196_random, chr8_gl000197_random, chr9_gl000198
_random, chr9_gl000199_random, chr9_gl000200_random, chr9_gl000201_random, chr11_gl000202_random,
chr17_ctg5_hap1, chr17_gl000203_random, chr17_gl000204_random, chr17_gl000205_random, chr17_gl0002
06_random, chr18_gl000207_random, chr19_gl000208_random, chr19_gl000209_random, chr21_gl000210_ran
dom, chrUn_gl000211, chrUn_gl000212, chrUn_gl000213, chrUn_gl000214, chrUn_gl000215, chrUn_gl00021
6, chrUn_gl000217, chrUn_gl000218, chrUn_gl000219, chrUn_gl000220, chrUn_gl000221, chrUn_gl000222,
chrUn_gl000223, chrUn_gl000224, chrUn_gl000225, chrUn_gl000226, chrUn_gl000227, chrUn_gl000228, ch
rUn_gl000229, chrUn_gl000230, chrUn_gl000231, chrUn_gl000232, chrUn_gl000233, chrUn_gl000234, chrU
n_gl000235, chrUn_gl000236, chrUn_gl000237, chrUn_gl000238, chrUn_gl000239, chrUn_gl000240, chrUn_
gl000241, chrUn_gl000242, chrUn_gl000243, chrUn_gl000244, chrUn_gl000245, chrUn_gl000246, chrUn_gl
000247, chrUn_gl000248, chrUn_gl000249]
```

goutham atla @Geek_y 8d

The moment when GATK complains about contigs order #Bioinformatics #GATK #NGS pic.twitter.com/OxVAb5ZyWH

- **Part 2 - Gene Models**

1. GeneModels : RefSeq, GENCODE
2. Comparison between the 2 gene models
3. We all live in a “NM_ world”
4. Which one to choose?

RefSeq Geneset

RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

1. Widely used gene set produced by the NCBI,
2. Has significant manually annotated content, but much less than GENCODE (~45% of transcripts are listed as MODEL),
3. Transcripts are named as:
   a. NM: Manually curated, protein-coding transcripts,
   b. NR: Non-coding transcrips,
   c. XM: Predicted protein-coding models.
4. ongoing curation by NCBI staff and collaborators, with reviewed records indicated

GENCODE Geneset

1. Goal : create reference gene annotations for the ENCODE project,
2. Comprehensive +++ (e.g. include pseudogenes, lncRNAs, short RNAs, protein-coding transcripts),
3. Extensive manual annotation by the HAVANA group, as well as computational annotation.
4. ~ 93.4% of the annotations involve manual annotation
5. Under constant validation by many groups in the consortium.
6. Default annotation set used by the Ensembl project.

HUMAN

GENCODE 30 (08.04.19)

GENCODE vs RefSeq Genesets

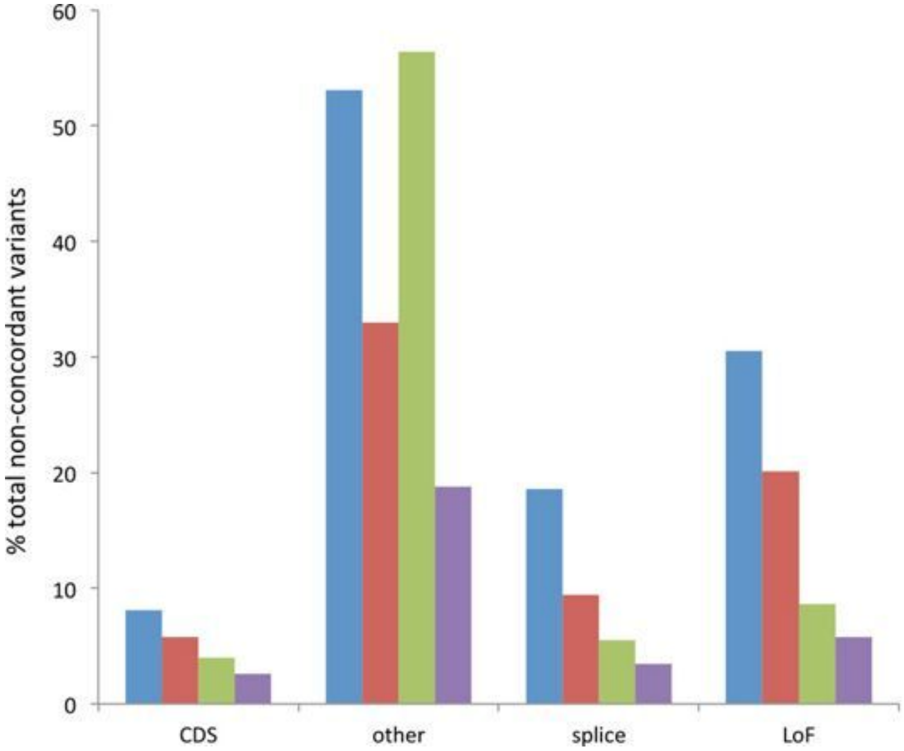
Category	GENCODE	RefSeq
PURPOSE	Enhancing and extending the annotation of all evidence-based gene features in the human genome at a high accuracy	Providing a comprehensive, integrated, non-redundant, well-annotated set of sequences (genomic, transcript and protein).
ANNOTATION The process of finding and designating locations of individual genes and other features on raw DNA sequences	**Primary transcriptomic data aligned to the reference genome to determine transcript structure and CDSs.** **+** Manual annotation : use of datasets that capture TSS and transcript 3' ends, epigenetic and transcription factor binding data as well as cross-species conservation	**Well-supported and biologically valid transcripts** reviewed by RefSeq curators at the NCBI. **RefSeq transcripts are annotated independently of the genome and based upon the mRNA sequence alone.** Curated transcripts aligned to the genome sequence and combined with additional computational models
SEQUENCE	GENCODE sequences always match the genome reference assembly.	RefSeq sequences don't necessarily match the genome reference assembly.

Impact of Gene Model on variant annotation

Larger source of difference between consequence predictions : Unique variants

Proportion of discordant calls :

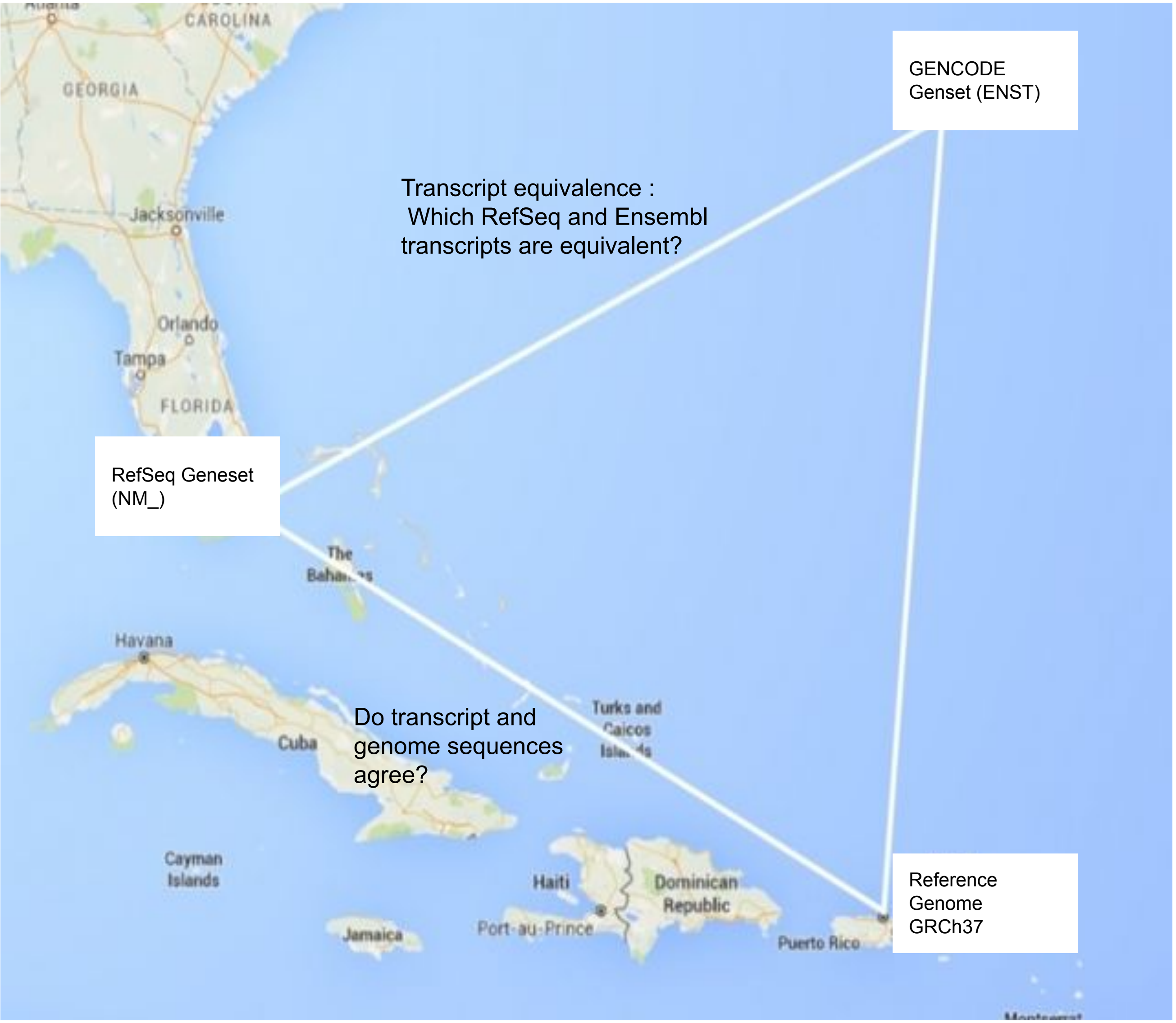
Dataset	GENCODE Comprehensive vs RefSeq NXR
1000 Genomes (WGS + WES)	3.1 %
ESP (WES only)	1.7 %



- CDS variants show high (>90%) concordance in all conditions
- 'Other' variants show high discordance (up to 56%).
- Approximately 30% of LoF variant calls are in conflict.

Frankish *et al.* 2015

We all live in a “NM_ world”



GENCODE Genset (ENST)

Transcript equivalence :
Which RefSeq and Ensembl transcripts are equivalent?

RefSeq Geneset (NM_)

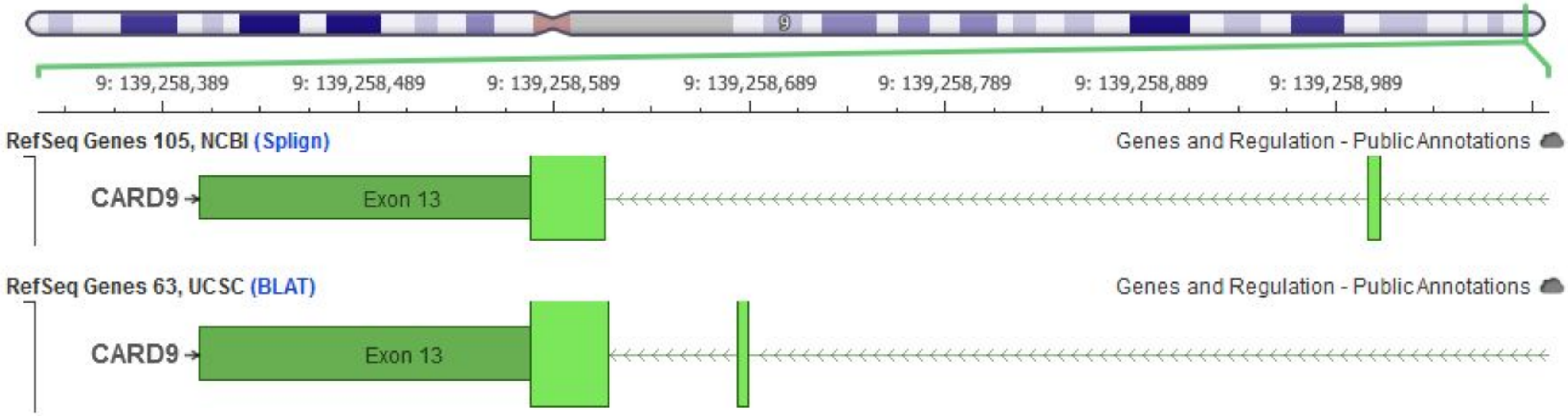
Do transcript and genome sequences agree?

Reference Genome GRCh37

From *The Clinical Significance of Transcript Alignment Discrepancies* presented by Reece Hart at Human Variome Project Meeting 2014, Paris

We all live in a “NM_ world”

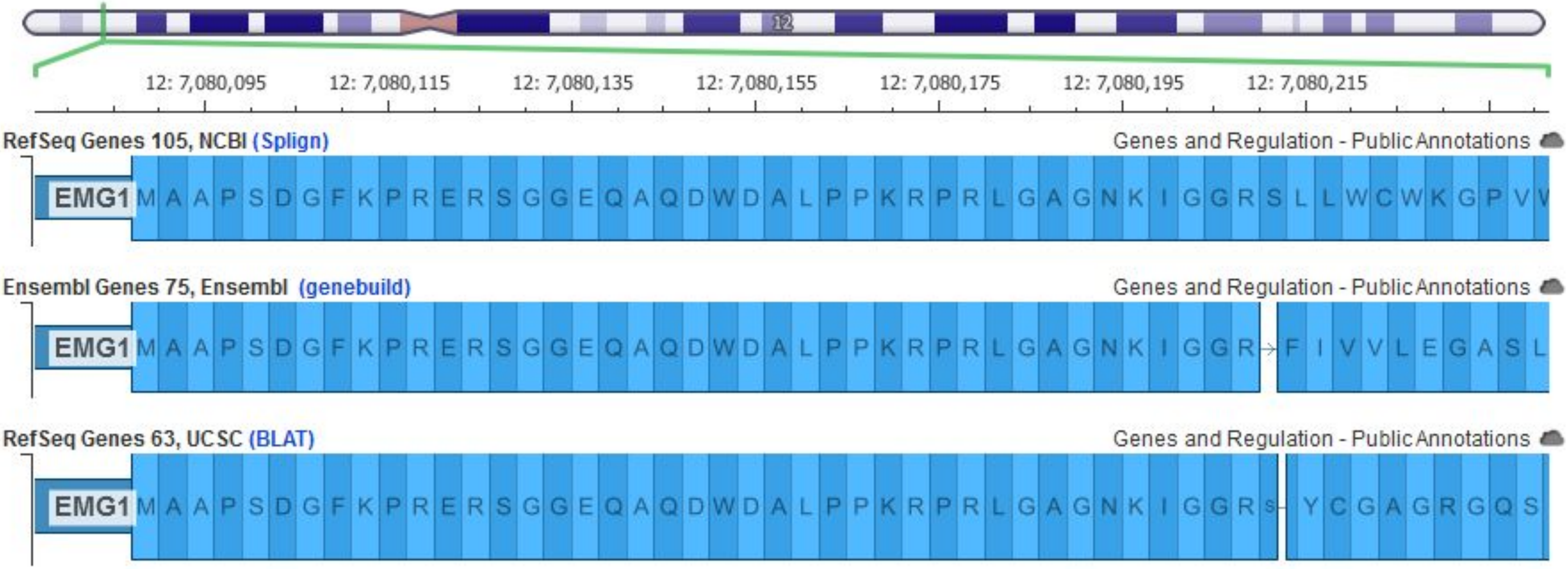
When transcript alignment discrepancies lead to discordant exon coordinates


9
9: 139,258,389
9: 139,258,489
9: 139,258,589
9: 139,258,689
9: 139,258,789
9: 139,258,889
9: 139,258,989
RefSeq Genes 105, NCBI (Splign)
Genes and Regulation - Public Annotations
CARD9
Exon 13
RefSeq Genes 63, UCSC (BLAT)
Genes and Regulation - Public Annotations
CARD9
Exon 13

https://blog.goldenhelix.com/using-the-grch38-reference-assembly-for-clinical-interpretation-in-vsclinical-webcast-qa/
https://blog.goldenhelix.com/refseq-genes-updated-to-ncbi-provided-alignments-and-why-you-care/

We all live in a “NM_ world”

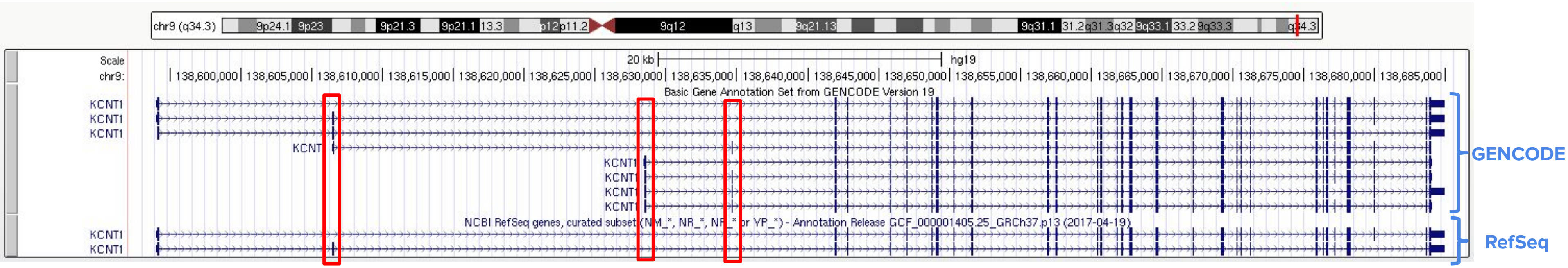
“gap” between the NM_006331 transcript’s RNA sequence and the human genomic sequence.


12
12: 7,080,095
12: 7,080,115
12: 7,080,135
12: 7,080,155
12: 7,080,175
12: 7,080,195
12: 7,080,215
RefSeq Genes 105, NCBI (Splign)
Genes and Regulation - Public Annotations
EMG1 M A A P S D G F K P R E R S G G E Q A Q D W D A L P P K R P R L G A G N K I G G R S L L W C W K G P V V
Ensembl Genes 75, Ensembl (genebuild)
Genes and Regulation - Public Annotations
EMG1 M A A P S D G F K P R E R S G G E Q A Q D W D A L P P K R P R L G A G N K I G G R F I V V L E G A S L
RefSeq Genes 63, UCSC (BLAT)
Genes and Regulation - Public Annotations
EMG1 M A A P S D G F K P R E R S G G E Q A Q D W D A L P P K R P R L G A G N K I G G R s Y C G A G R G Q S

https://blog.goldenhelix.com/refseq-genes-updated-to-ncbi-provided-alignments-and-why-you-care/

So which transcript set should we choose?

Available GENCODE and RefSeq transcripts for the KCNT1 gene

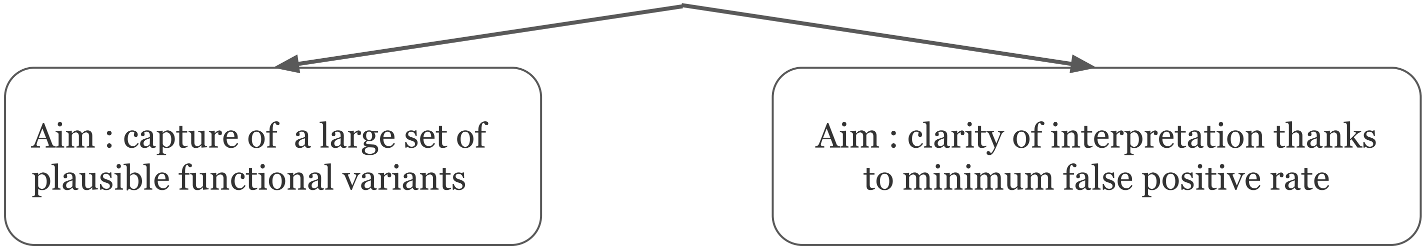


Novel transcription start site exons and novel internal exons not present in RefSeq.

Not an isolated case.

No Best Choice…

Dichotomy at the heart of variant annotation:



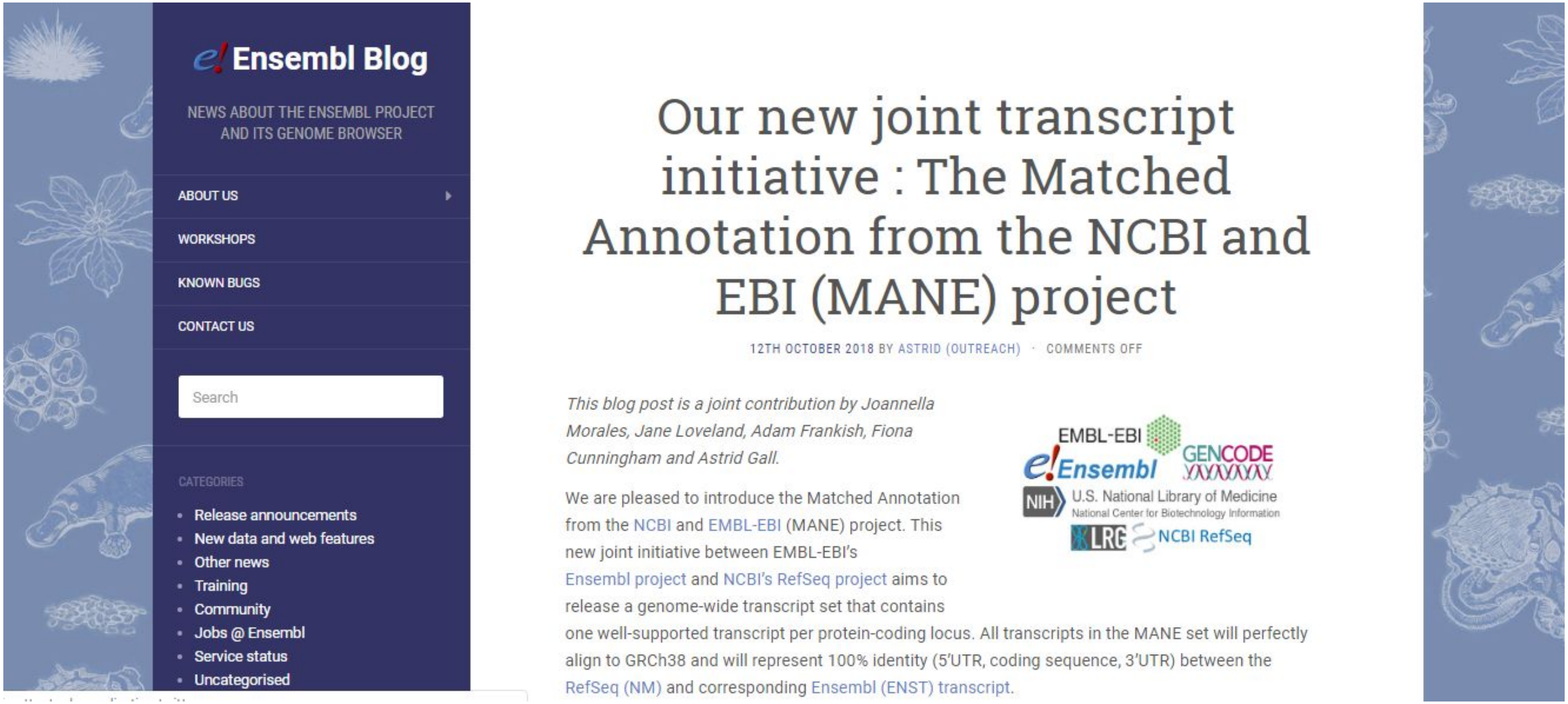
Aim : capture of a large set of plausible functional variants

Aim : clarity of interpretation thanks to minimum false positive rate

“When choosing an annotation database, researchers should keep in mind that no database is perfect and some gene annotations might be inaccurate or entirely wrong.”

Assessing the impact of human genome annotation choice on RNA-seq expression estimates. Wu *et al.* 2013
BMC Bioinformatics. 2013;14(Suppl 11):S8. doi: 10.1186/1471-2105-14-S11-S8.

There is still hope...

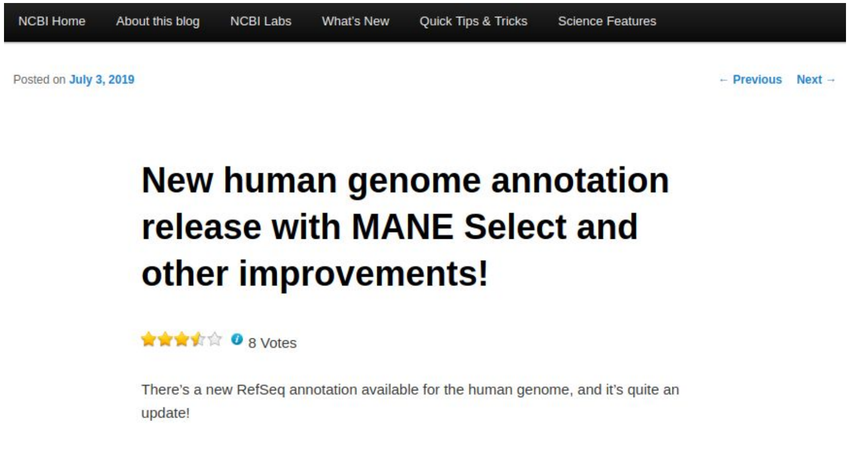


Our new joint transcript initiative : The Matched Annotation from the NCBI and EBI (MANE) project

12TH OCTOBER 2018 BY ASTRID (OUTREACH) · COMMENTS OFF

This blog post is a joint contribution by Joannella Morales, Jane Loveland, Adam Frankish, Fiona Cunningham and Astrid Gall.

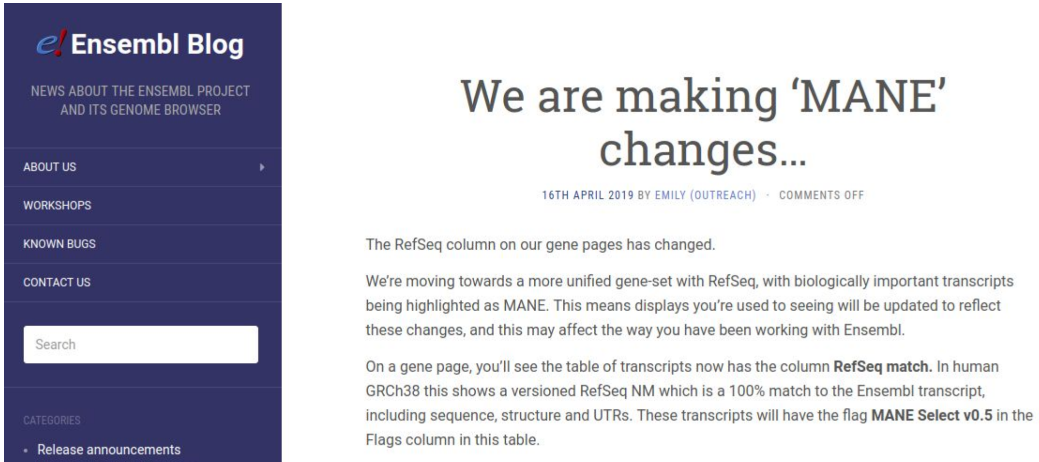
We are pleased to introduce the Matched Annotation from the NCBI and EMBL-EBI (MANE) project. This new joint initiative between EMBL-EBI's Ensembl project and NCBI's RefSeq project aims to release a genome-wide transcript set that contains one well-supported transcript per protein-coding locus. All transcripts in the MANE set will perfectly align to GRCh38 and will represent 100% identity (5'UTR, coding sequence, 3'UTR) between the RefSeq (NM) and corresponding Ensembl (ENST) transcript.



Posted on July 3, 2019

New human genome annotation release with MANE Select and other improvements!

There's a new RefSeq annotation available for the human genome, and it's quite an update!



We are making 'MANE' changes...

16TH APRIL 2019 BY EMILY (OUTREACH) · COMMENTS OFF

The RefSeq column on our gene pages has changed.

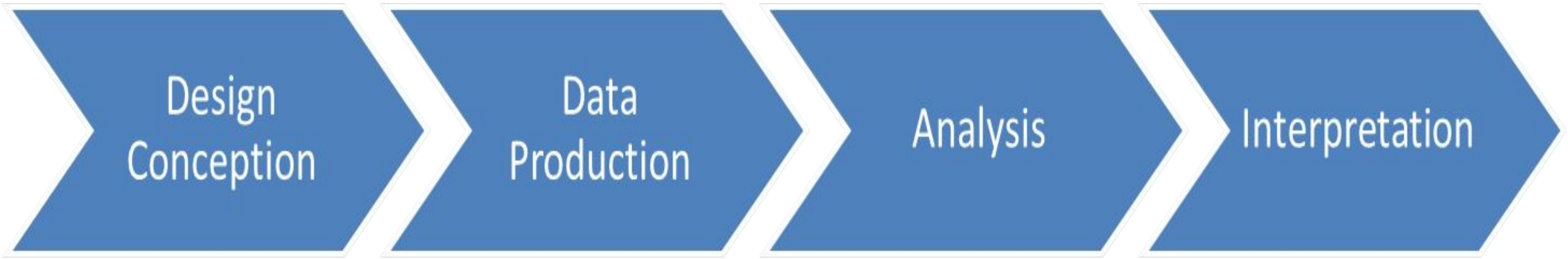
We're moving towards a more unified gene-set with RefSeq, with biologically important transcripts being highlighted as MANE. This means displays you're used to seeing will be updated to reflect these changes, and this may affect the way you have been working with Ensembl.

On a gene page, you'll see the table of transcripts now has the column **RefSeq match.** In human GRCh38 this shows a versioned RefSeq NM which is a 100% match to the Ensembl transcript, including sequence, structure and UTRs. These transcripts will have the flag **MANE Select v0.5** in the Flags column in this table.

- **CONCLUSION**

Take home message :

Reference Genome and Gene Model do impact your NGS Workflow !



Design Conception

Data Production

Analysis

Interpretation

Which Gene Model and reference genome were used to select targeted regions in your design ?

Which reference genome was used to analyze your data?

Which Gene Model was used to annotate your variants ?

Keep in mind :

- transcript equivalence
- strength and weakness of both Reference Genome and Gene Model you rely on