



The University of Manchester

The ever changing human reference genome sequence

Dr Peter Causey-Freeman

Lecturer in Healthcare Sciences (Clinical-Bioinformatics)



The University of Manchester

Who am I?

Dr Peter Causey-Freeman

Lecturer in Healthcare Sciences (Clinical Bioinformatics)



Affiliations:

[Division of Informatics, Imaging & Data Sciences \(L5\)](#)

/ Division of Informatics, Imaging & Data Sciences

Links: [LinkedIn profile](#)

Email: peter.causey-freeman@manchester.ac.uk

Phone: +44(0) 161 275 5731

[Full contact details](#)

ORCID: 0000-0002-5838-5404



[View graph
of relations](#)



[https://www.research.manchester.ac.uk/portal/en/researchers/peter-causeyfreeman\(f3b9aafa-80b7-48fd-be97-04f835bd0363\).html](https://www.research.manchester.ac.uk/portal/en/researchers/peter-causeyfreeman(f3b9aafa-80b7-48fd-be97-04f835bd0363).html)

What I do

VariantValidator

Accurate validation, mapping and formatting of sequence variants using HGVS nomenclature.



What We Do

We validate HGVS sequence variation descriptions, accurately mapping between transcript and genomic variants. We also automate conversion of genomic (VCF) sequence variation descriptions into the HGVS format and vice-versa.

VariantValidator auto-corrects your mistakes if it can and helps you correct your own if it can't. We provide a range of tools to meet your needs including batch processing, a VCF file converter and API access.

Powered By

VariantValidator
version 1.0.2.dev3+gec89acd
vv_hgvs
version 1.2.5.vv1
UTA
release uta_20180821
SeqRepo
release 2018-08-21

Validator

Validate your variant descriptions using HGVS nomenclature.

[Try It Out!](#)

Batch Validator

Validate multiple variant descriptions at once.

[Try It Out!](#)

VCF to HGVS

Convert VCF files to validated HGVS variants.

[Try It Out!](#)

Gene to Transcript

Identify all transcripts from a gene symbol.

[Try It Out!](#)

- Diverse global user community
- User feedback evolved the software to meet the community need
- Automatic lift-over between GRCh37 and 38
- 5 years of headaches!
 - Technical issues
 - Helping users

A brief history of the human genome sequence













And why GRCh37 like an out-of-date map on a Sat-Nav

The human genome project 1990 - 2003

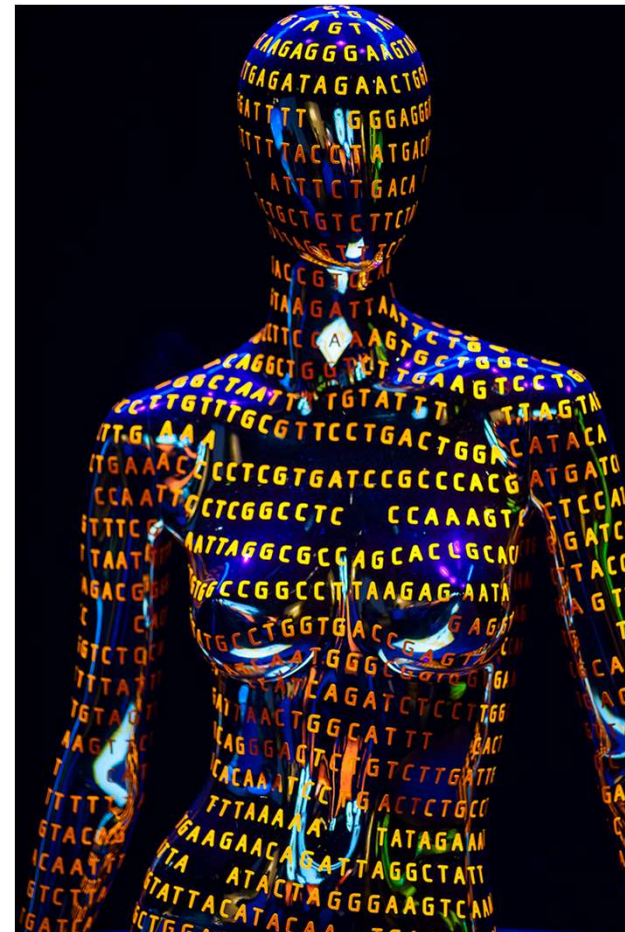
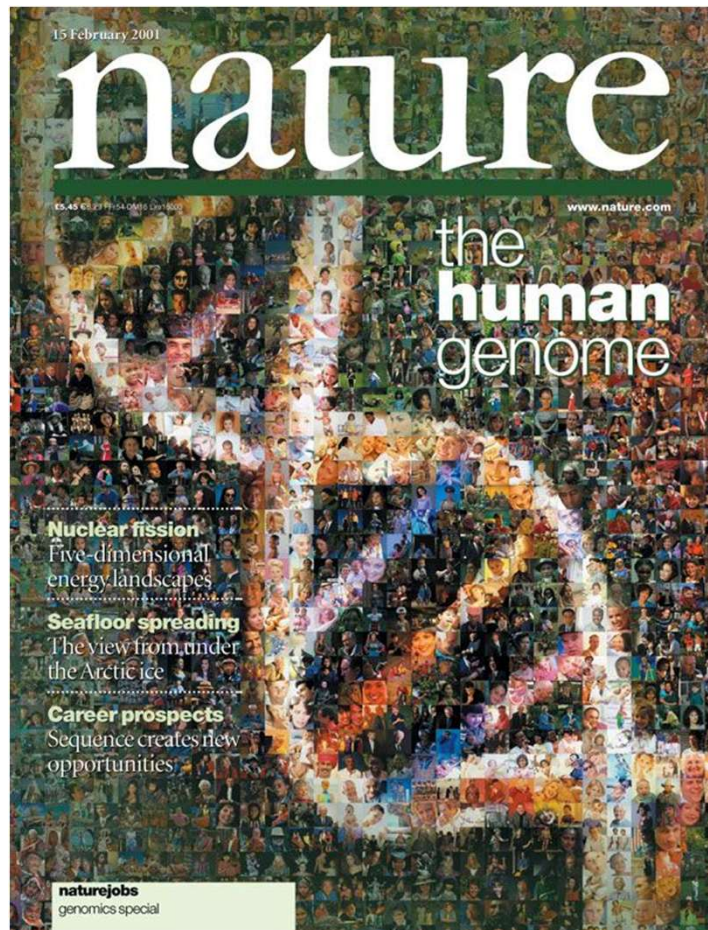
- 13-year-long, publicly funded project
- OBJECTIVE – Determine the sequence of the entire **euchromatic** human genome
 - Lightly packed chromatin
 - Enriched in genes
 - ~92% of the human genome
- Major funding from the US NIH and DOE
 - Budgeted \$3BN US
 - Estimated final cost ~\$5BN

An international sequencing project

The institutions, companies, and laboratories in Human Genome Program are listed below, according to [NIH](#).^[4]

No.	Nation	Name	Affiliation
1		The Whitehead Institute/MIT Center for Genome Research	Massachusetts Institute of Technology
2		The Wellcome Trust Sanger Institute	Wellcome Trust
3		Washington University School of Medicine Genome Sequencing Center	Washington University in St. Louis
4		United States DOE Joint Genome Institute	United States Department of Energy
5		Baylor College of Medicine Human Genome Sequencing Center	Baylor College of Medicine
6		RIKEN Genomic Sciences Center	Riken
7		Genoscope and CNRS UMR-8030	French Alternative Energies and Atomic Energy Commission
8		GTC Sequencing Center	Genome Therapeutics Corporation , whose sequencing division is acquired by ABI
9		Department of Genome Analysis	Fritz Lipmann Institute  , name changed from Institute of Molecular Biotechnology
10		Beijing Genomics Institute/Human Genome Center	Chinese Academy of Sciences
11		Multimegabase Sequencing Center	Institute for Systems Biology
12		Stanford Genome Technology Center	Stanford University
13		Stanford Human Genome Center and Department of Genetics	Stanford University School of Medicine
14		University of Washington Genome Center	University of Washington
15		Department of Molecular Biology	Keio University School of Medicine
16		University of Texas Southwestern Medical Center at Dallas	University of Texas
17		University of Oklahoma's Advanced Center for Genome Technology	Dept. of Chemistry and Biochemistry, University of Oklahoma
18		Max Planck Institute for Molecular Genetics	Max Planck Society
19		Lita Annenberg Hazen Genome Center	Cold Spring Harbor Laboratory
20		GBF/German Research Centre for Biotechnology	Reorganized and renamed to Helmholtz Center for Infection Research 

2001- the draft genome sequence released



Key analysis findings

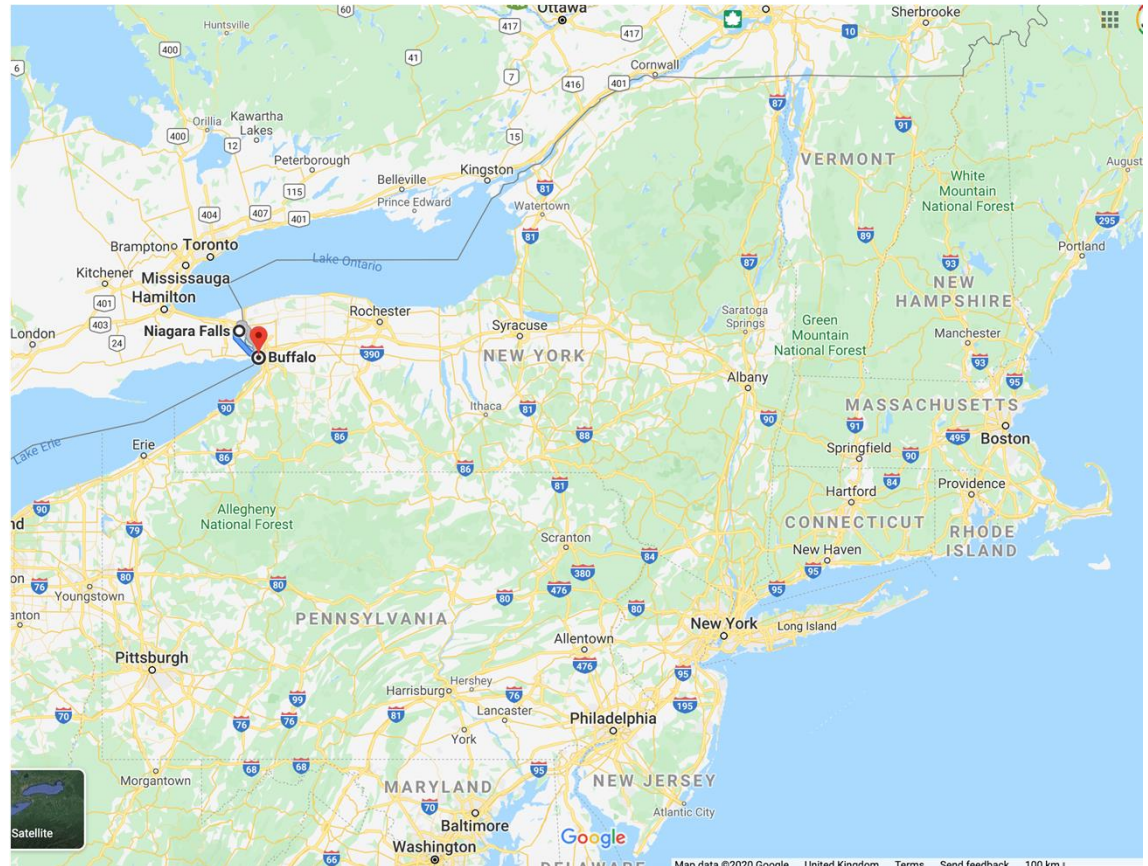
- ~22,300 protein coding genes
 - ~25% of the genome
 - ~1% is exon
- Significantly more segmental duplications than expected
 - Nearly identical, repeated sections of DNA
 - Functional gene copies (e.g. insulin)
 - Pseudogenes
- Huge number of Short Interspersed Nuclear Elements
 - Primarily *Alu* elements
 - ~15% of the genome

The average sequence of the average human (*right????*)

- DNA collected from a large number of donors (estimated ~30?)
 - blood (female)
 - sperm (male)
- Only a few processed into DNA libraries
 - insures anonymity
- Much of the sequence came from 3 donors
 - 2 female
 - 1 male

The average sequence of the average human (*right????*)

- Due to quality considerations, >70% of the sequence came from the 1 man
 - Mr RP11
 - From Buffalo, New York



Difference between the draft and final genome sequences

- Defined by coverage and accuracy
- 2001
 - ~90% genome coverage (euchromatic)
 - ~1/1,000 error rate
 - ~150,000 gaps (un-sequenced regions)
- 2003
 - ~99% genome coverage (euchromatic)
 - ~1/10,000 error rate
 - ~400 gaps

There are still gaps and errors

- GRCh37
 - ~300 gaps
 - ~550 genes with indel errors
- GRCh38
 - ~89 “unresolved” gaps as of June 2019 (GRCh38.p13)
 - A few hundred genes with indel errors (GRCh38.p1)
- GRCh39?
 - It’s coming!

Our map needs regular updates because roads keep changing

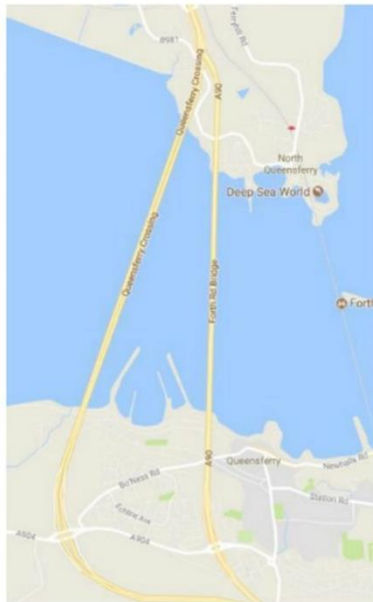
News / Scottish News

The Queensferry Crossing has now been added to Google Maps... but not some sat navs



by Ross Crae

🕒 August 30, 2017, 4:53 pm



The Queensferry Crossing is now visible on most maps (Google & Jane Barlow / PA Wire)

**B. MAIN
SCULPTORS**



Order before the
end of February
and all headstones
and markers will
come with
**FREE
LETTERING**

[Read More](#)

The Queensferry Crossing opened to the public for the first time one year ago today

Potential hazards of an out-of-date map map?



How the genome reference sequence evolves

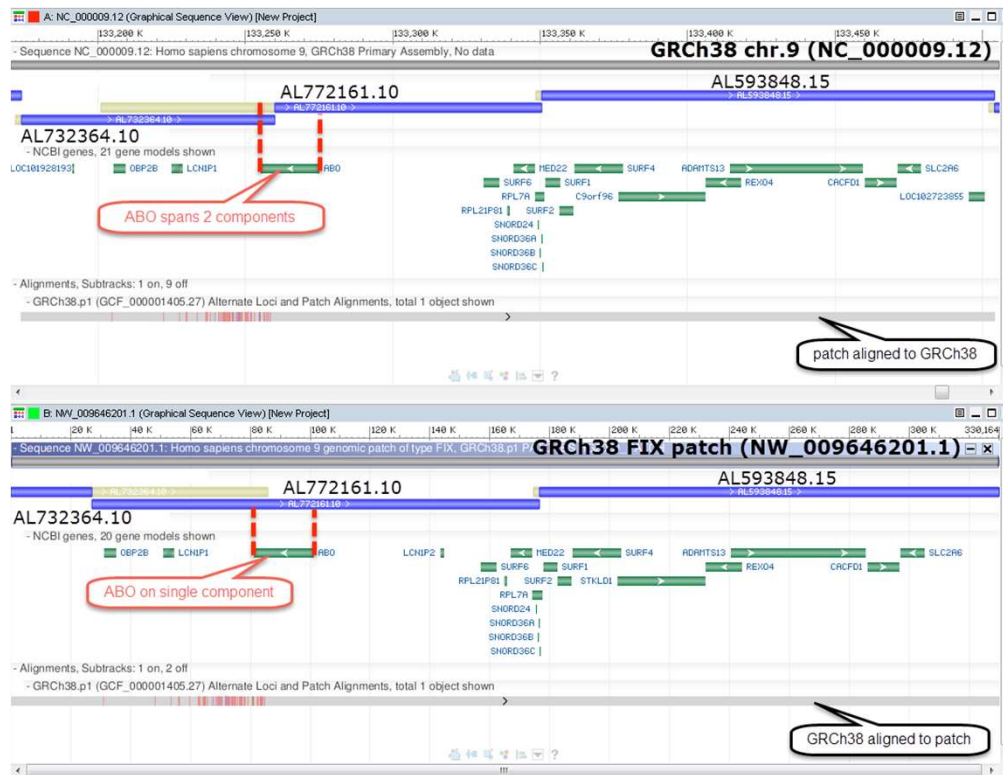
A basic overview of patches and Alt Loci

What is a patch

Add information to the assembly without disrupting chromosome coordinates or sequences

- Reference sequences with unique identifiers
- Assembly units used to create/correct genomic reference sequences
- Given Chromosome context by alignment to the current genome build

Fix patches correct gaps or sequence errors



- Within the ABO gene, GRCh38 has an absent base
 - Incorrect switching between scaffolds
- Patch scaffold created that “fixes” the error

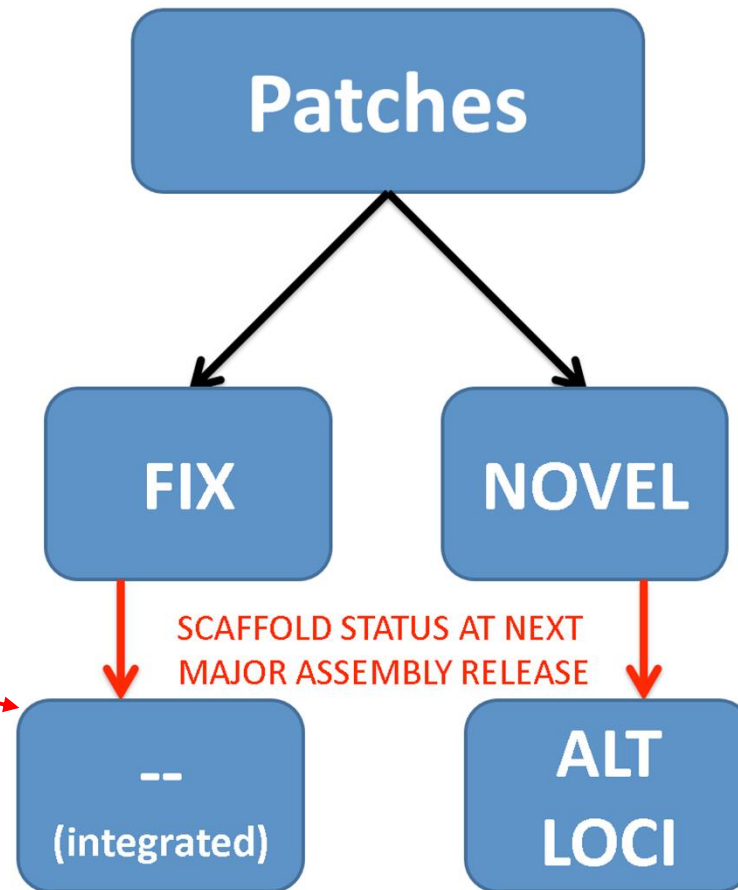
Novel patch

- An alternate “structure” for a chromosomal region, e.g. CYP206 duplication



How patches affect the genome build

- Coordinates and sequence of the primary assembly only change during major releases



- Patch reference sequences deprecated at the major release

Some genes only found on patches or ALT loci

- Some genes have missing exons in the GRCh38 primary assembly (e.g. SHANK3)
- Some genes are not represented in the GRCh38 primary assembly (e.g. HLA-DRB4)
- A few hundred genes like ABO still have indel errors in the GRCh38 primary assembly
 - Many clinically relevant genes corrected GRCh37 > 38

Improvement GRCh37 > 38

- NC_000015.9:g.72105933del (GRCh37)
 - NR2E3
 - Retinal Dystrophy gene panels
 - Additional base in the GRCh37 chr15
 - gnomAD AF (linked via VV) ~99%

HGVS-compliant variant descriptions

Type	Variant Description	Link to Reference sequence Record
Transcript (:c.)	NM_014249.3:c.946_949=	NM_014249.3
RefSeq Gene (:g.)	NG_009113.1:g.8034_8037=	NG_009113.1
Protein (:p.)	NP_055064.1:p.(Asp316=)	NP_055064.1
Protein (:p.)	NP_055064.1:p.(D316=)	NP_055064.1

<https://variantvalidator.org/>



Hypothetical issue: have you checked your data for -

- NC_000015.9:g.72105929= (GRCh37)
 - Extrapolate gnomAD AF - could be up to ~1%

HGVS-compliant variant descriptions

Type	Variant Description	Link to Reference sequence Record
Transcript (:c.)	NM_014249.3:c.951dup	NM_014249.3
RefSeq Gene (:g.)	NG_009113.1:g.8039dup	NG_009113.1
Protein (:p.)	NP_055064.1:p.(Thr318HisfsTer23)	NP_055064.1
Protein (:p.)	NP_055064.1:p.(T318Hfs*23)	NP_055064.1

- NC_000015.10:g.71813592dup (GRCh38)
 - Variation would likely have been identified



Reference genome sequences and genes

My take on a few issues waiting in the wings!



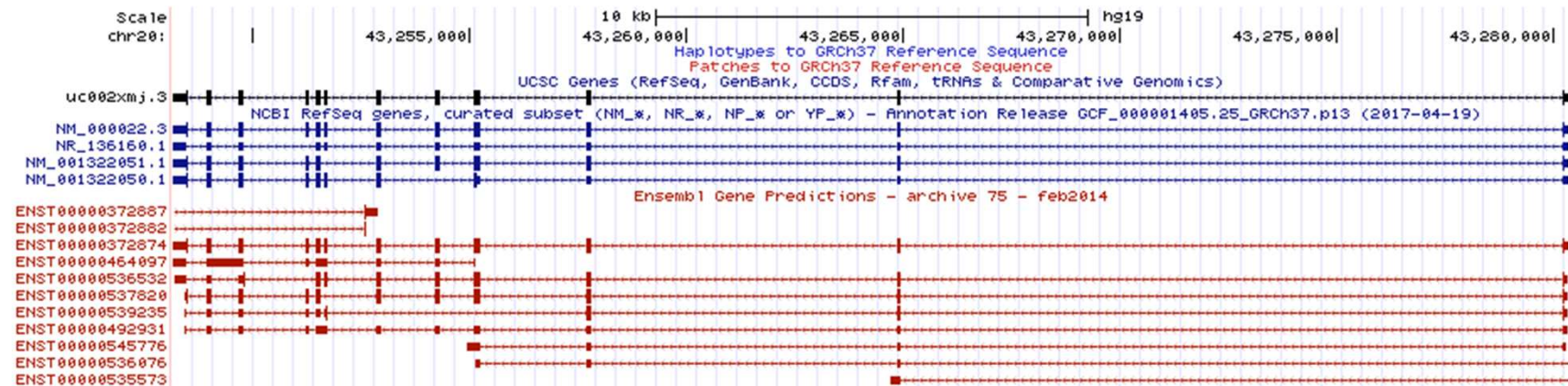
Gene annotations (The basics)

Alignment file from source

- Chromosome IDs
- Transcript ID
- Exon to genome aln. data
- Orientation
- Other metadata (e.g. coding vs non-coding, CDS coordinates)

Reference sequence record

- Transcript sequence
- Metadata
 - CDS start/end
 - Translation ID / Sequence
 - Publications
 - Features



Variability dependent on the source

RefSeq vs Ensembl approach

RefSeq – Transcript to Genome

- Transcript and genomic sequences are independent
 - May disagree
- The same version of the transcript can usually align to GRCh37 and GRCh38
 - Not always
- In rare cases, silent re-annotation occurs (e.g. alignment software update)
 - Exon boundaries shift without changes to the reference sequence
 - No incrementing in the version number
- **HGVS now recommend recording the genomic variant**

RefSeq vs Ensembl approach

Ensembl – Genome to Transcript

- Transcript and genomic sequences not independent
 - Always
- New transcript version created for each genome build
 - Requires update to the transcript variant
 - More difficult to lift-over between genome builds
- The major annotation is fixed
 - Exon boundaries cannot shift without incrementing

Updating your GRCh37 data

- Best case – Transcript aligned to both 37 and 38

HGVS-compliant variant descriptions

Type	Variant Description	Link to Reference sequence Record
Transcript (:c.)	NM_000088.3:c.589G>T	NM_000088.3
Transcript (:c.)	LRG_1t1:c.589G>T	LRG_1
RefSeq Gene (:g.)	NG_007400.1:g.8638G>T	NG_007400.1
LRG (:g.)	LRG_1:g.8638G>T	LRG_1
Protein (:p.)	NP_000079.2(LRG_1p1):p.(Gly197Cys)	NP_000079.2
Protein (:p.)	NP_000079.2:p.(G197C)	NP_000079.2

Genomic Variants

Variant Description	VCF Description	Link to GenBank
NC_000017.11:g.50198002C>A	GRCh38:17:50198002:C:A	NC_000017.11
NC_000017.10:g.48275363C>A	GRCh37:17:48275363:C:A	NC_000017.10

Updating your GRCh37 data

- Transcript does not aligned to both 37 and 38

HGVS-compliant variant descriptions

Type	Variant Description	Link to Reference sequence Record
Transcript (:c.)	NM_000022.2:c.534A>G	NM_000022.2
Transcript (:c.)	LRG_16t1:c.534A>G	LRG_16
RefSeq Gene (:g.)	NG_007385.1:g.32462A>G	NG_007385.1
LRG (:g.)	LRG_16:g.32462A>G	LRG_16
Protein (:p.)	NP_000013.2(LRG_16p1):p.(Val178=)	NP_000013.2
Protein (:p.)	NP_000013.2:p.(V178=)	NP_000013.2

Genomic Variants

Variant Description	VCF Description	Link to GenBank
NC_000020.10:g.43252915T>C	GRCh37:20:43252915:T:C	NC_000020.10

Updating your GRCh37 data

- Project to genome and select updated transcript

Select	Transcript Accession	Gene Accession	Latest Version
<input type="checkbox"/>	NM_000022.2	NG_007385.1	✗
<input checked="" type="checkbox"/>	NM_000022.3		✓
<input type="checkbox"/>	NM_001322050.1		✓
<input type="checkbox"/>	NM_001322051.1		✓
<input type="checkbox"/>	NR_136160.1		✓

HGVS-compliant variant descriptions

Type	Variant Description	Link to Reference sequence Record
Transcript (:c.)	NM_000022.3:c.534A>G	NM_000022.3
Protein (:p.)	NP_000013.2(LRG_16p1):p.(Val178=)	NP_000013.2
Protein (:p.)	NP_000013.2:p.(V178=)	NP_000013.2

Genomic Variants

Variant Description	VCF Description	Link to GenBank
NC_000020.11:g.44624274T>C	GRCh38:20:44624274:T:C	NC_000020.11
NC_000020.10:g.43252915T>C	GRCh37:20:43252915:T:C	NC_000020.10



LRG?

- Static reference sequences with static annotation
 - Update requires the creation of an entirely new LRG
 - Defeats the intended “scope” of the project
 - What happens when GRCh39 is released?
- Could RefSeq align NM_000022.2 (LRG_16t1) to GRCh38?
 - Cannot expect them to do this – replaced by .3
 - MANE project > NM_000022.4
 - Some existing LRGs refractory to Ensembl transcripts

What's in a name?

- HGVS are stressing the use of complete and unique reference sequence identifiers (Why?)
 - <http://varnomen.hgvs.org/bg-material/consultation/svd-wg008/>
- Which of these statements is correct?
 1. chrM (GRCh37) == chrM (hg19)
 2. chr1 (GRCh37) == chr1 (GRCh38)
 3. HSCHR6_MHC_MCF_CTG1 (GRCh37) == HSCHR6_MHC_MCF_CTG1 (GRCh38)
 4. HSCHR6_MHC_MCF_CTG1 (GRCh37) == chr6_ssto_hap7 (hg19)

What's in a name?

- **Common Mistake (Alamut Visual)**
 - chr6(GRCh37):g.135726089del
- **Easily becomes**
 - chr6:g.135726089del
- Use a unique reference sequence ID
 - NC_000006.11:g.135726089del
- And take care with your software
 - NC_000006.11:g.135726092del

Think ahead?

- **GRCh39 intends to represent different populations in the assembly**
- How many chr1(GRCh39) will there be?
 - If >1 , chr1(GRCh39) is not unique
- Use unique identifiers
 - GRCh38 - NC_000001.11
 - GRCh39 - NC_000001.12, NC_000001.13 etc. (?)

Concluding remarks

- Reference sequences are rapidly evolving
 - Push to GRCh39
 - MANE
- Data are becoming more complex
- Humans like data to be simple
- For the most part, we will likely get away with it during GRCh37 > 38
- Simple (boring) changes can be implemented now and might save us a headache when we are back here talking about GRCh38 > 39