Building the Reference Assembly

Deanna M. Church
Senior Director of Genomics and Content
Personalis, Inc

The Genome Reference Consortium consists of:

The Wellcome Trust Sanger Institute

The Genome Institute at Washington University

The European Bioinformatics Institute

The National Center for Biotechnology Information

Valerie Schneider, NCBI

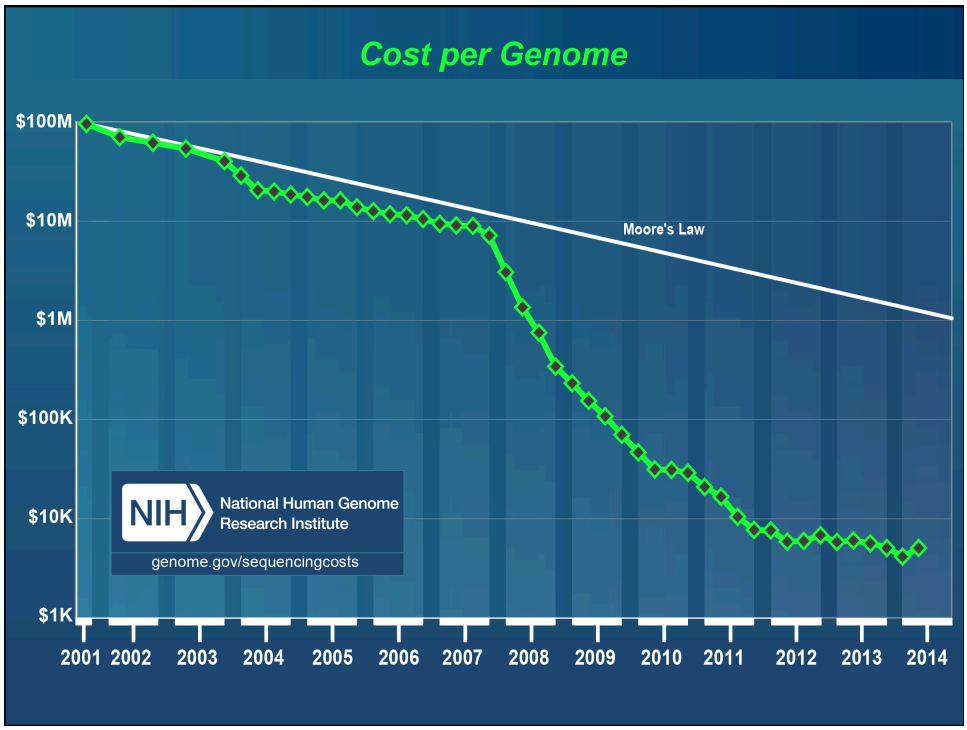
http://genomereference.org

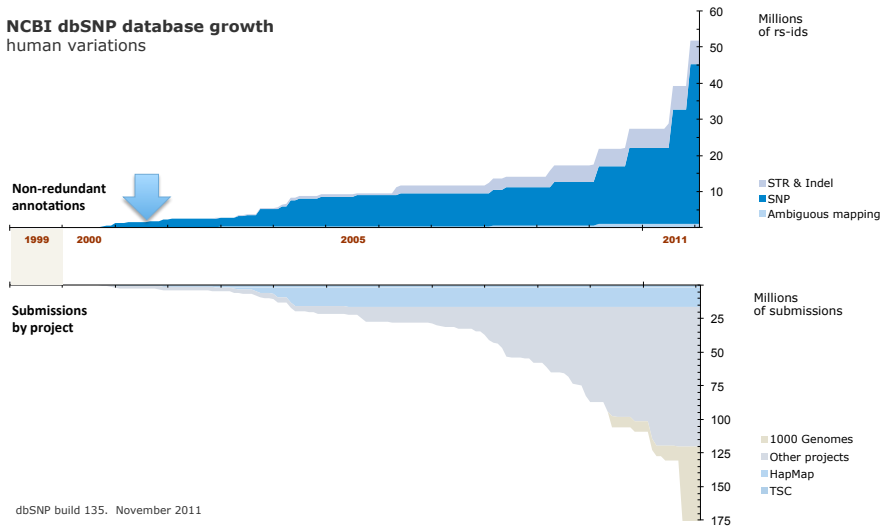
HGP Goals

Area	Goals 1993-98	Status as of Oct. 1998	Goals 1998-2003
Genetic map	Average 2- to 5-cM resolution	1 cM map published Sept. 1994	Completed
Gene identification	Develop technology	30,000 ESTs mapped	Full-length cDNAs
Functional analysis	Not a goal	-	Develop genomic-scale technologies
Model organisms	*E. coli*: complete sequence	Published Sept. 1997	-
	Yeast: complete sequence	Released Apr. 1996	-
	C. elegans: most of sequence	80% complete	Complete Dec. 1998
	Drosophila: begin sequencing	9% done	Sequence by 2002
	Mouse: map 10,000 STSs	12,000 STSs mapped	Develop extensive genomic resources
			Lay basis for finishing sequence by 2005 Produce working draft before 2005

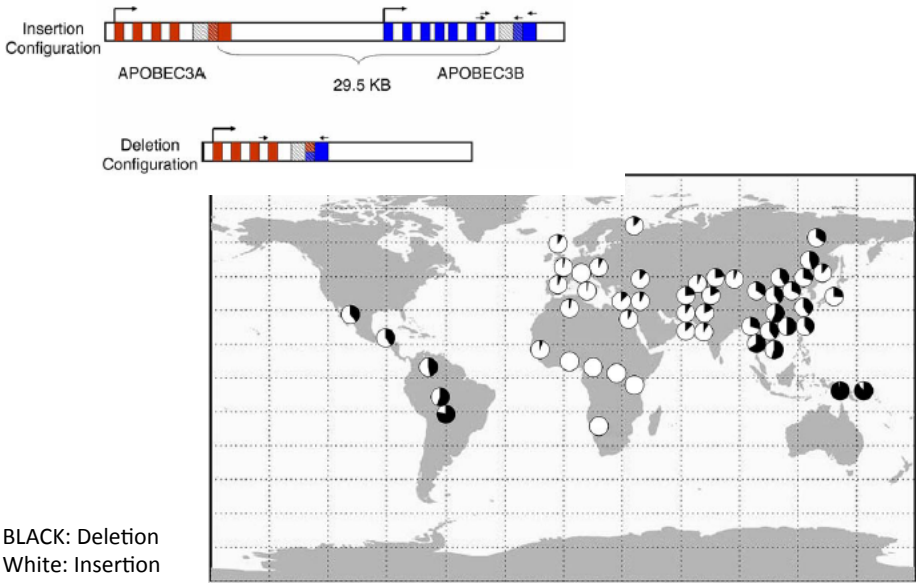
Throughput: 500 Mb/year
Cost: < $0.25 per base
Variation: 100,000 SNPs mapped

Collins FS et al, 1998



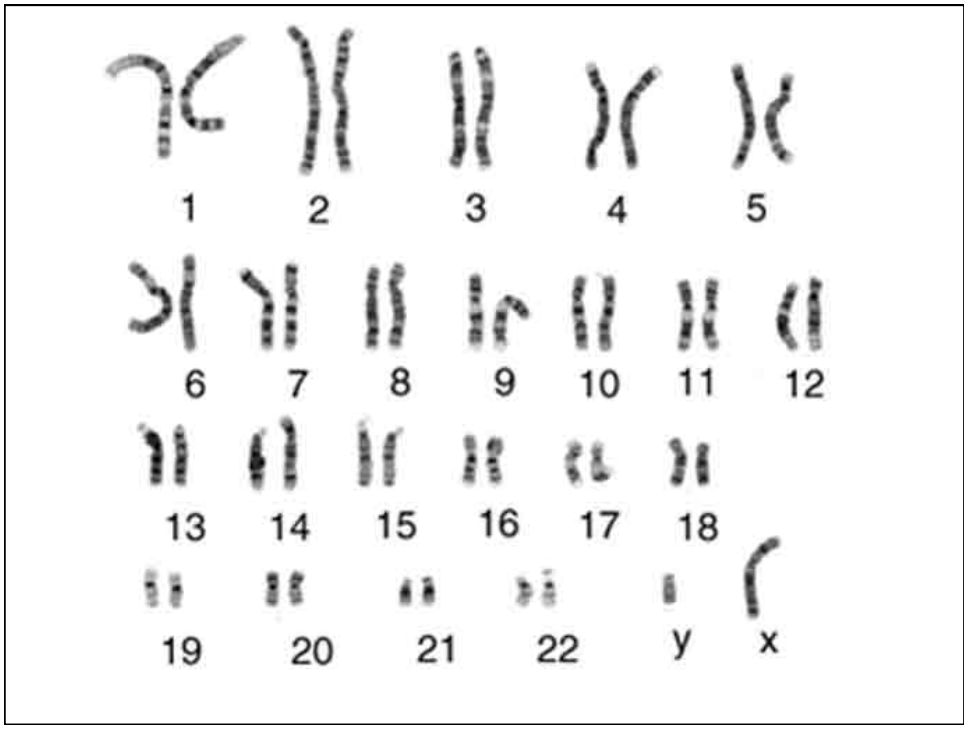


Steve Sherry, NCBI



BLACK: Deletion
White: Insertion

Kidd et al, 2007 APOBEC cluster



>gi|224514618|ref|NT_077402.2| Homo sapiens chromosome 1 genomic contig, GRCh37.p10 Primary Assembly

TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAAC
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTAA
CCCTAACCCCTAACCCTAACCCTAACCCTAACCCTAACCTAACCCTAACCCTAACCCTAACCCTAACCCT
AACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAACCCCAACCCCAACCCTAACCCCTAACCCTAACCCTAACC
CTACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCTAACCCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAACCCTAACCCTAACCCTCGCGGTACCCTCAGCCGGCCCG
CCCGCCCGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAGAGTACCACCGAAATCTGTGCAGAGGAC
AACGCAGCTCCGCCCTCGCGGTGCTCTCCGGGTCTGTGCTGAGGAGAACGCAACTCCGCCGTTGCAAAGG
CGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCG
GCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGAGGCGCGCCGCGCCGGCGCAGGCGCAG
AGAGGCGCGCCGCGCCGGCGCAGGCGCAGACACATGCTAGCGCGTCGGGGTGGAGGCGTGGCGCAGGCGC
AGAGAGGCGCGCCGCGCCGGCGCAGGCGCAGAGACACATGCTACCGCGTCCAGGGGTGGAGGCGTGGCGC
AGGCGCAGAGAGGCGCACCGCGCCGGCGCAGGCGCAGAGACACATGCTAGCGCGTCCAGGGGTGGAGGCG
TGGCGCAGGCGCAGAGACGCAAGCCTACGGGCGGGGGTTGGGGGGGCGTGTGTTGCAGGAGCAAAGTCGC
ACGGCGCCGGGCTGGGGCGGGGGGAGGGTGGCGCCGTGCACGCGCAGAAACTCACGTCACGGTGGCGCGG
CGCAGAGACGGGTAGAACCTCAGTAATCCGAAAAGCCGGGATCGACCGCCCCTTGCTTGCAGCCGGGCAC
TACAGGACCCGCTTGCTCACGGTGCTGTGCCAGGGCGCCCCCTGCTGGCGACTAGGGCAACTGCAGGGCT
CTCTTGCTTAGAGTGGTGGCCAGCGCCCCCTGCTGGCGCCGGGGCACTGCAGGGCCCTCTTGCTTACTGT
ATAGTGGTGGCACGCCGCCTGCTGGCAGCTAGGGACATTGCAGGGTCCTCTTGCTCAAGGTGTAGTGGCA
GCACGCCCACCTGCTGGCAGCTGGGGACACTGCCGGGCCCTCTTGCTCCAACAGTACTGGCGGATTATAG
GGAAACACCCGGAGCATATGCTGTTTGGTCTCAGTAGACTCCTAAATATGGGATTCCTGGGTTTAAAAGT
AAAAAATAAATATGTTTAATTTGTGAACTGATTACCATCAGAATTGTACTGTTCTGTATCCCACCAGCAA
TGTCTAGGAATGCCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCC
TGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGTGTATTTTTTTAAATTTCCACTGATGATT
TTGCTGCATGGCCGGTGTTGAGAATGACTGCGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATGTAG
TTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTT
CTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTGTCTCTTAGCCCAGACTTCCCGTGTCCTTTCCACC
GGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTGGTCTTCATCTGCAGGTGTCTGACTTCCAGCAACTG
CTGGCCTGTGCCAGGGTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAG
TGGGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATACCACAACCAGGCATAG
GGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTTCTCTCACAACCTAGGCCAGTAAGTAGTGCTT
GTGCTCATCTCCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCTG

Reference assembly history

PERSPECTIVE

Against a Whole-Genome Shotgun

Philip Green[1]

Department of Molecular Biotechnology, University of Washington, Seattle, Washington 98195

PERSPECTIVE

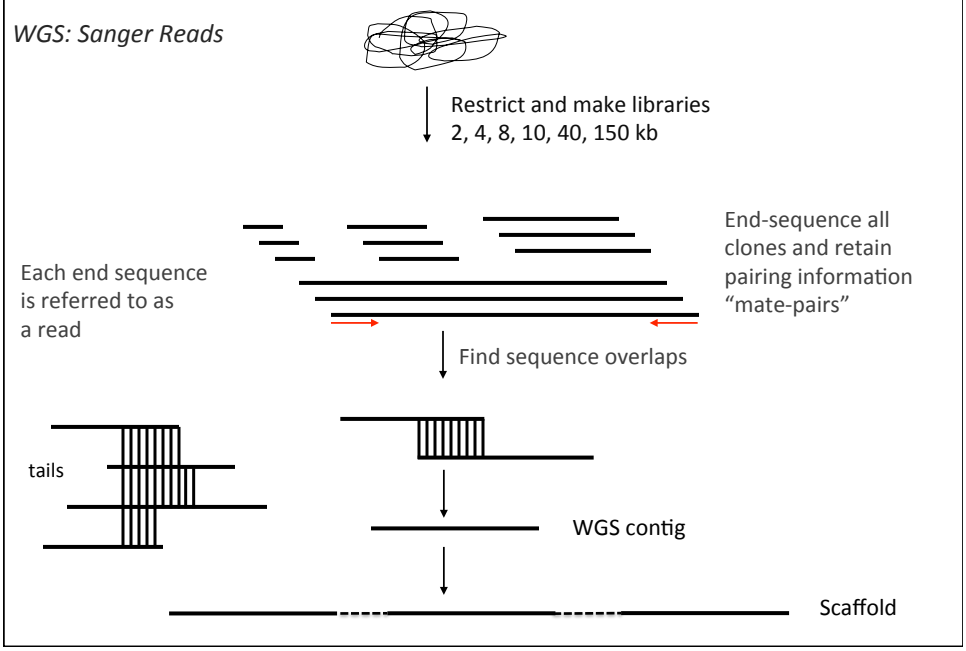
Human Whole-Genome Shotgun Sequencing

James L. Weber[1,3] **and Eugene W. Myers**[2]

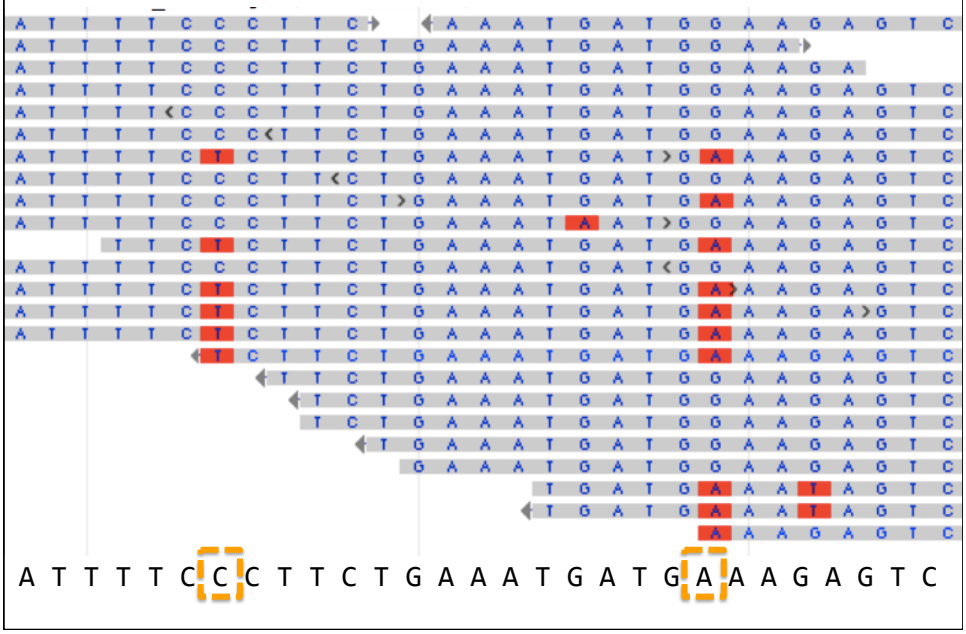
[1]Center for Medical Genetics, Marshfield Medical Research Foundation, Marshfield, Wisconsin 54449;
[2]Department of Computer Science, University of Arizona, Tucson, Arizona 85721

Genome Research, May, 1997

Reference assembly history

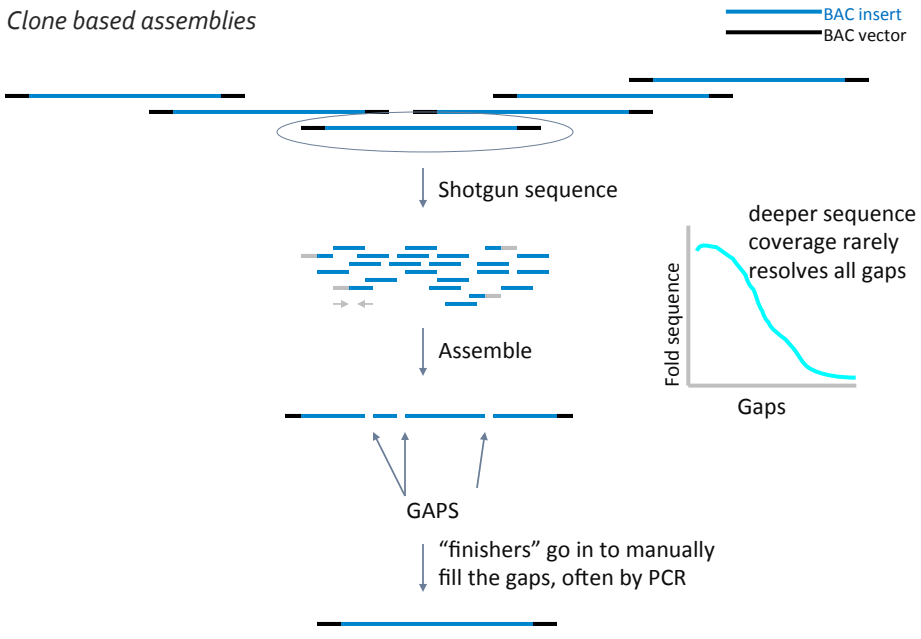


Reference assembly history



Reference assembly history

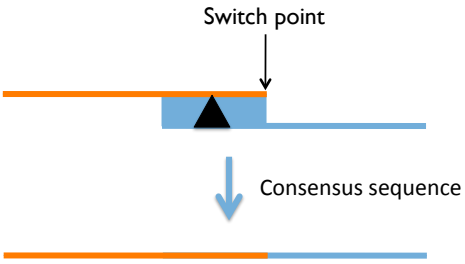
Clone based assemblies

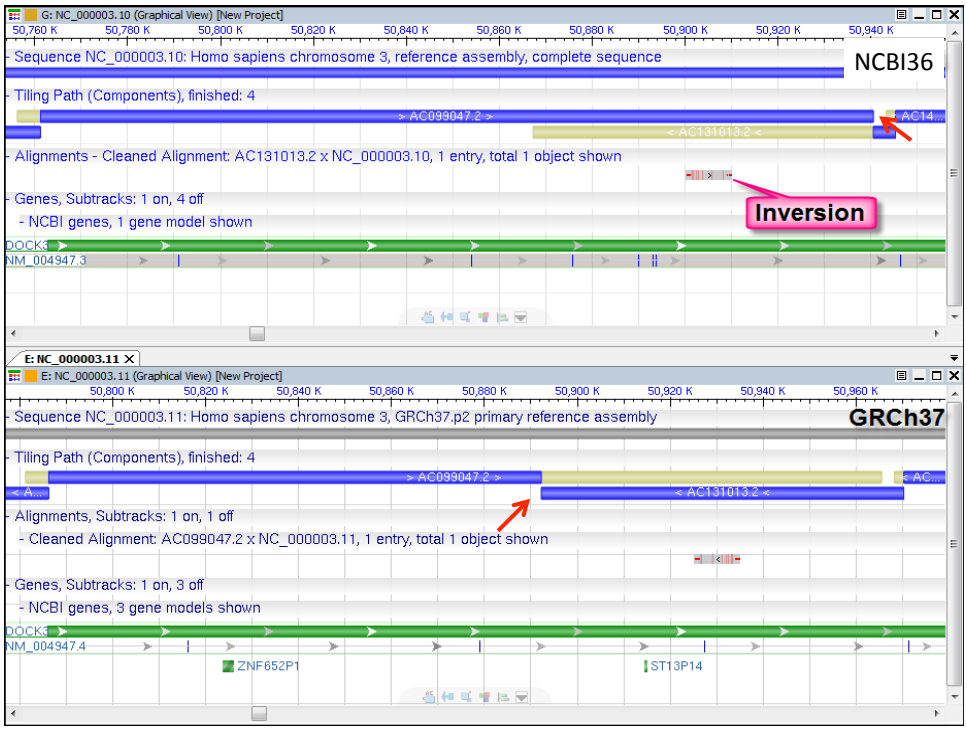


Reference assembly history

Build sequence contigs based on contigs defined in TPF.

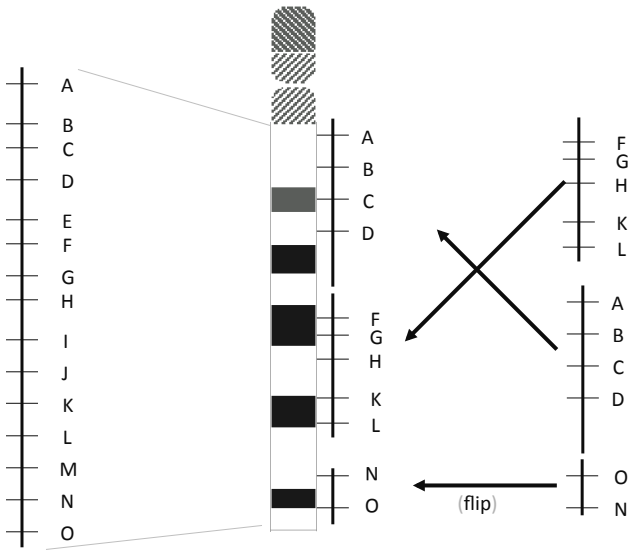
- Check for orientation consistencies
- Select switch points
- Instantiate sequence for further analysis





Reference assembly history

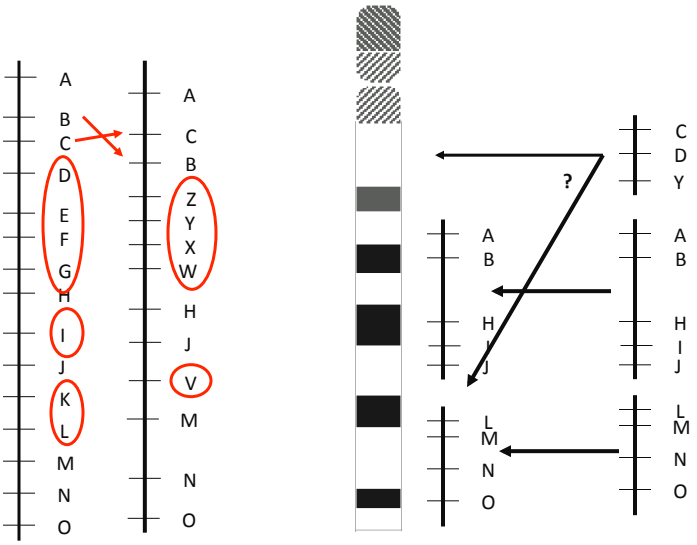
Ideally…

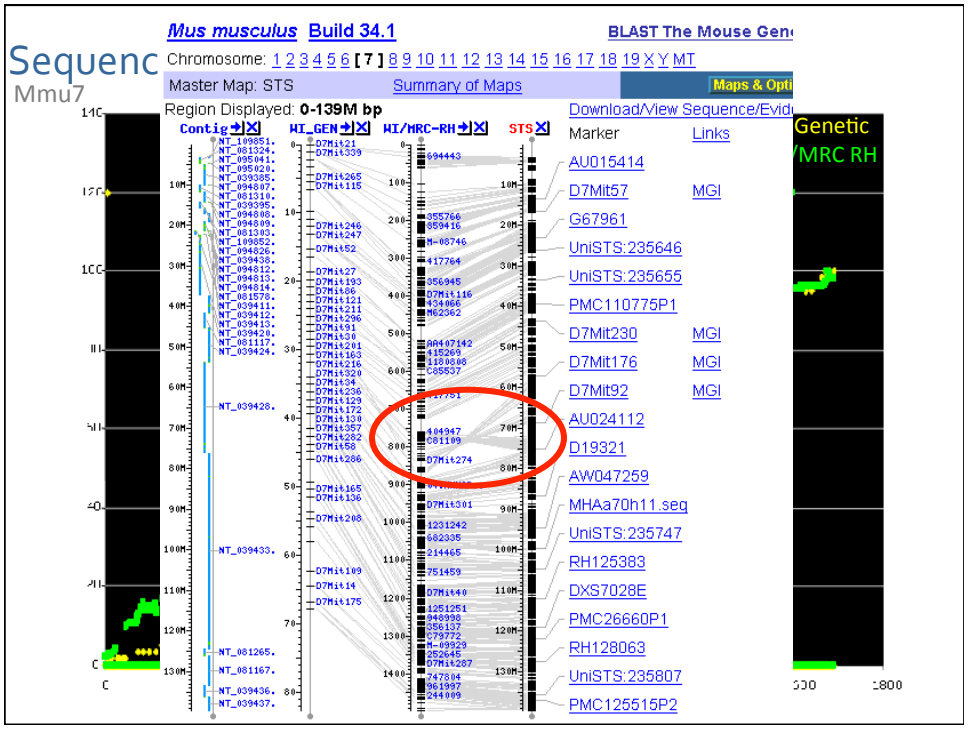


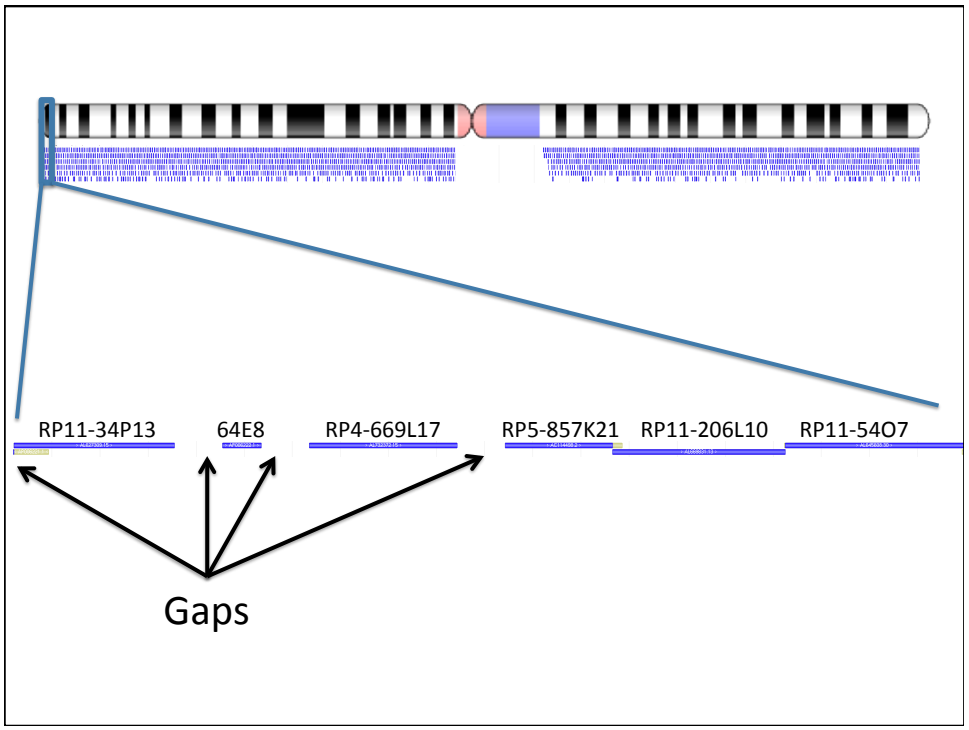
Non-sequence based Map

Reference assembly history

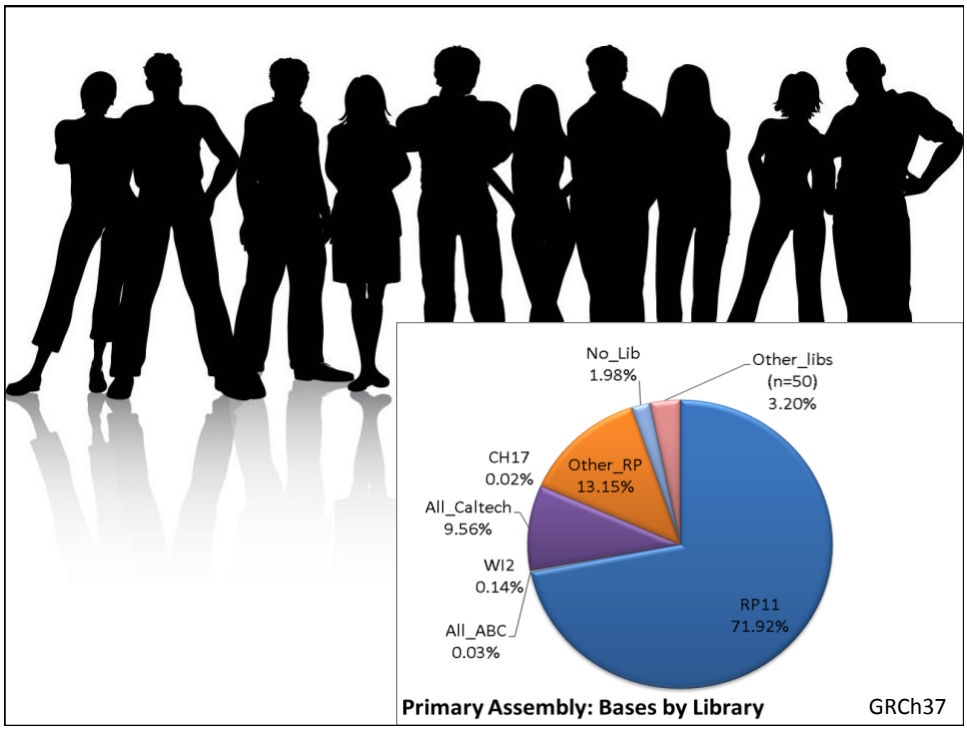
More like…

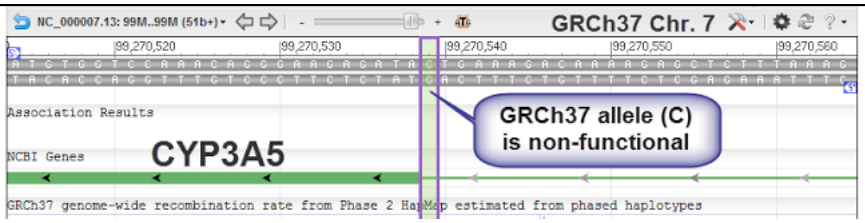



Mus musculus Build 34.1
BLAST The Mouse Gen
Sequenc
Mmu7
Chromosome: 1 2 3 4 5 6 [7] 8 9 10 11 12 13 14 15 16 17 18 19 X Y MT
Master Map: STS
Summary of Maps
Maps & Opti
Region Displayed: 0-139M bp
Download/View Sequence/Evid
Contig
MI_GEN
MI/MRC-RH
STS
Marker
Links
AU015414
D7Mit57 MGI
G67961
UniSTS:235646
UniSTS:235655
PMC110775P1
D7Mit230 MGI
D7Mit176 MGI
D7Mit92 MGI
AU024112
D19321
AW047259
MHAa70h11.seq
UniSTS:235747
RH125383
DXS7028E
PMC26660P1
RH128063
UniSTS:235807
PMC125515P2
Genetic
/MRC RH


RP11-34P13
64E8
RP4-669L17
RP5-857K21
RP11-206L10
RP11-54O7
Gaps

An assembly is a MODEL of the genome

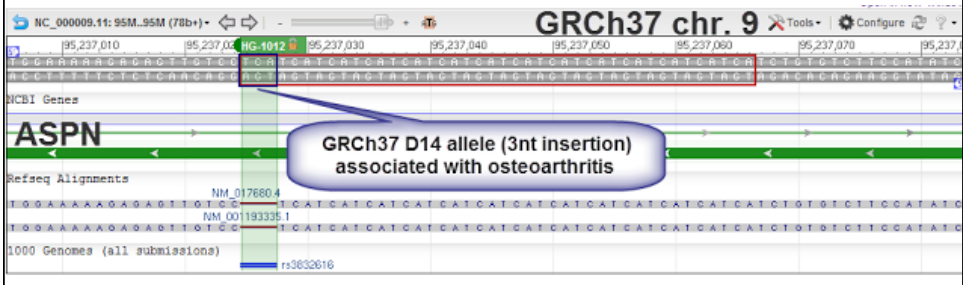




Issue Report for HG-1321

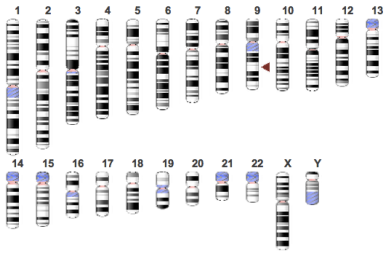
Category:	Variation
Affects version(s):	GRCh37
Report type:	RefSeq Report
Description:	The reference may be representing the non-protein-coding allele of GeneID:1577 (CYP3A5)
Last updated:	2012-12-03
Status:	Open
Experiment type:	na

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/issue_detail.cgi?id=HG-1321



Issue Report for HG-1012

Category:	Variation
Affects version(s):	GRCh37
Report type:	RefSeq Report
Description:	The reference genome is representing the D14 (less common) allele of NM_017680.4
Last updated:	2013-03-01
Status:	Awaiting External Info
Experiment type:	na



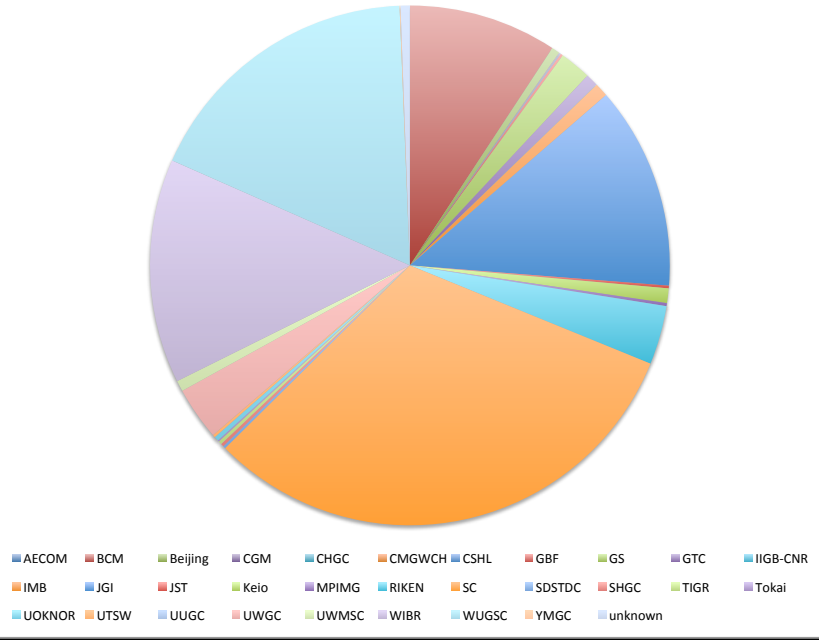
Assembly Information

Select a placement below to display it in the Sequence Viewer.

- GRCh37.p12 chr9:95,118,957-95,271,297 (View Regions: Ensembl | NCBI | UCSC)
- NCBI36 chr9:94,158,778-94,311,118 (View Regions: Ensembl | NCBI | UCSC)

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/issue_detail.cgi?id=HG-1012

Center sequence distribution: NCBI36



Finding the data

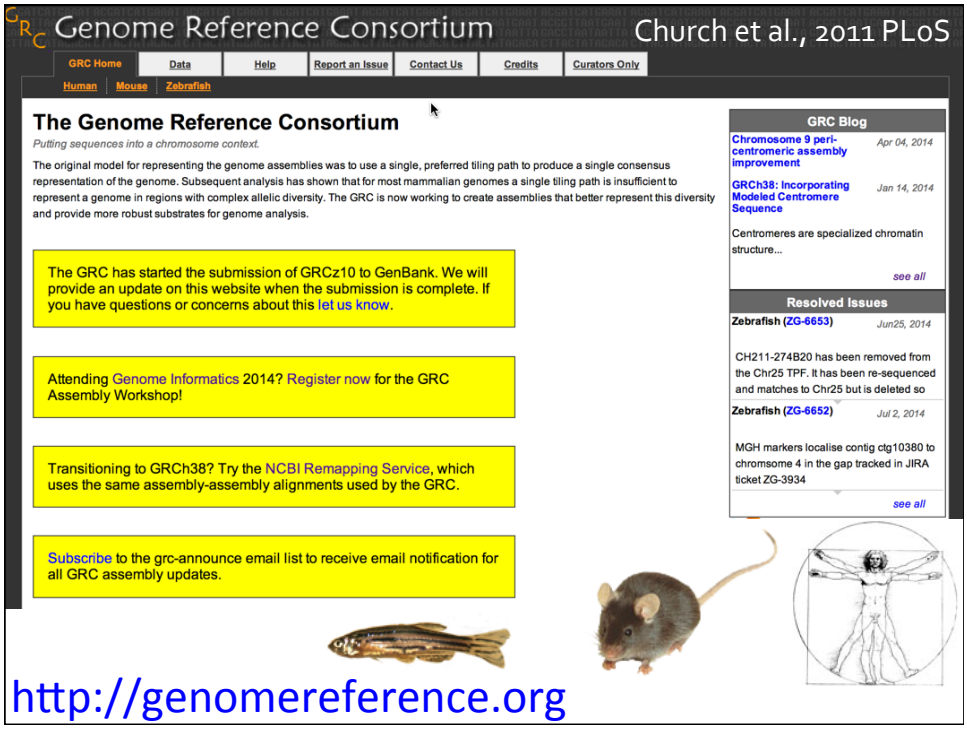
ends of each insert. Centres differed in the fluorescent labels employed and in the degree to which they used dye-primers or dye-terminators. The sequence detectors included both slab gel- and capillary-based devices. Detailed protocols are available on the web sites of many of the individual centres (URLs can be found at http://www.nhgri.nih.gov/genome_hub.html). The extent of

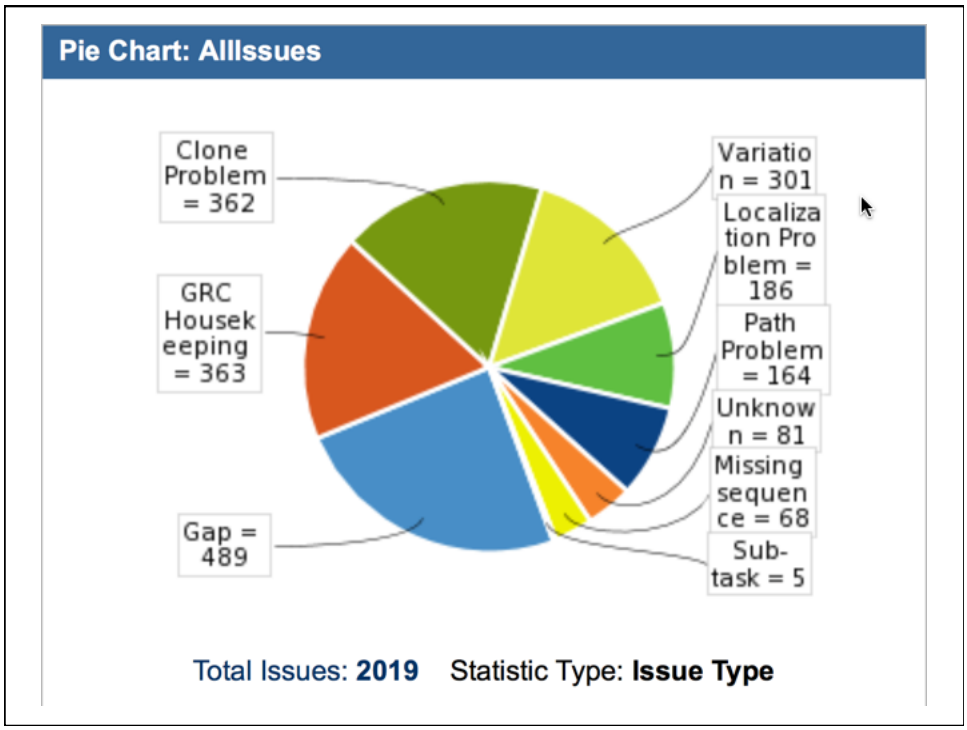
404 - Not Found

The link you followed or typed is either incomplete, outdated, contains disallowed characters, or we've made a mistake!

- Please try our *Powered-by-Google* search engine above (your best bet).
- Navigate to what you want directly from the Home page.
- Please send us feedback if you think there is a problem with our site.

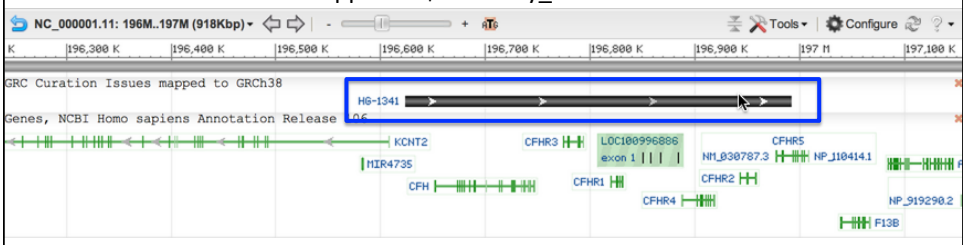
Our apologies for any inconvenience!


Genome Reference Consortium
Church et al., 2011 PLoS
GRC Home
Data
Help
Report an Issue
Contact Us
Credits
Curators Only
Human
Mouse
Zebrafish
The Genome Reference Consortium
Putting sequences into a chromosome context.
The original model for representing the genome assemblies was to use a single, preferred tiling path to produce a single consensus representation of the genome. Subsequent analysis has shown that for most mammalian genomes a single tiling path is insufficient to represent a genome in regions with complex allelic diversity. The GRC is now working to create assemblies that better represent this diversity and provide more robust substrates for genome analysis.
The GRC has started the submission of GRCz10 to GenBank. We will provide an update on this website when the submission is complete. If you have questions or concerns about this let us know.
Attending Genome Informatics 2014? Register now for the GRC Assembly Workshop!
Transitioning to GRCh38? Try the NCBI Remapping Service, which uses the same assembly-assembly alignments used by the GRC.
Subscribe to the grc-announce email list to receive email notification for all GRC assembly updates.
GRC Blog
Chromosome 9 peri-centromeric assembly improvement
Apr 04, 2014
GRCh38: Incorporating Modeled Centromere Sequence
Jan 14, 2014
Centromeres are specialized chromatin structure...
see all
Resolved Issues
Zebrafish (ZG-6653)
Jun25, 2014
CH211-274B20 has been removed from the Chr25 TPF. It has been re-sequenced and matches to Chr25 but is deleted so
Zebrafish (ZG-6652)
Jul 2, 2014
MGH markers localise contig ctg10380 to chromsome 4 in the gap tracked in JIRA ticket ZG-3934
see all
http://genomereference.org

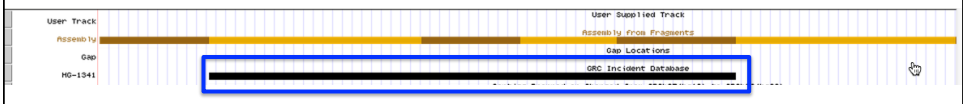

Pie Chart: AllIssues
Clone Problem = 362
GRC Housekeeping = 363
Gap = 489
Variation = 301
Localization Problem = 186
Path Problem = 164
Unknown = 81
Missing sequence = 68
Sub-task = 5
Total Issues: 2019 Statistic Type: Issue Type

Finding Curated Regions: Browsers

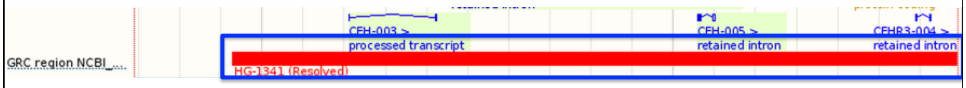
NCBI: GRC Curation Issues mapped to $assembly_name



UCSC: GRC Incident Database



Ensembl: GRC_region_NCBI37



Finding Curated Regions: GRC web browser

Issue Report for HG-1341

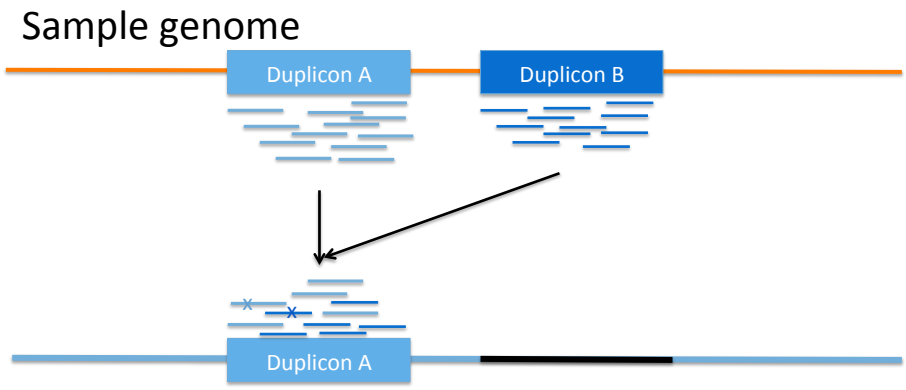
Category:	Unknown
Affects version(s):	GRCh37
Report type:	TPF Analysis
Description:	Clones have been selected for sequencing that align near the AL049744.8 - AL139418.9 region of Chr. 1. These may be used to improve the reference or to represent an alternate haplotype.
Last updated:	2013-07-18
Status:	Resolved
Experiment type:	Clone Sequencing
Resolution:	This CH17 pathway does not meet the criteria to replace the reference or for alt_loci creation.
Fix version(s):	GRCh38

Assembly Information

Select a placement below to display it in the Sequence Viewer.

- (•) GRCh38 chr1:196,623,707-196,992,672 (View Regions: NCBI | UCSC)
- ( ) GRCh37.p13 chr1:196,592,837-196,961,802 (View Regions: Ensembl | NCBI | UCSC)

Missing Sequence: Gaps



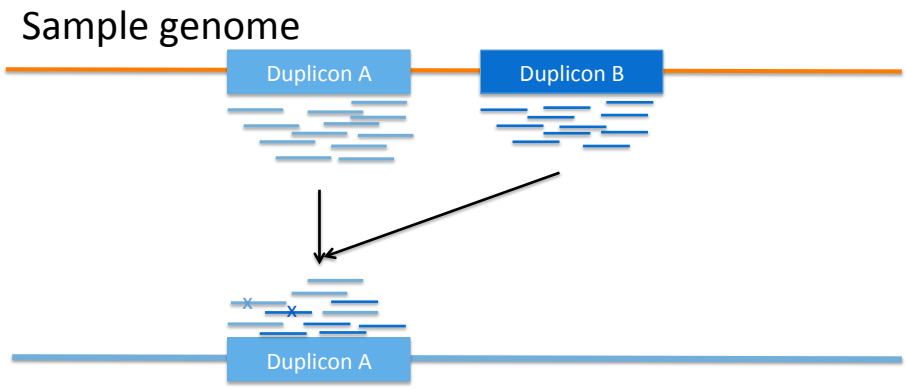
GRCh37

G>A (allelic difference – true variant)

G>C (paralogous sequence variant- false positive)

May or may not detect increased coverage depending on sequencing depth
and library quality (easier to find with new technologies than with old, low through technologies)

Missing sequence: collapse

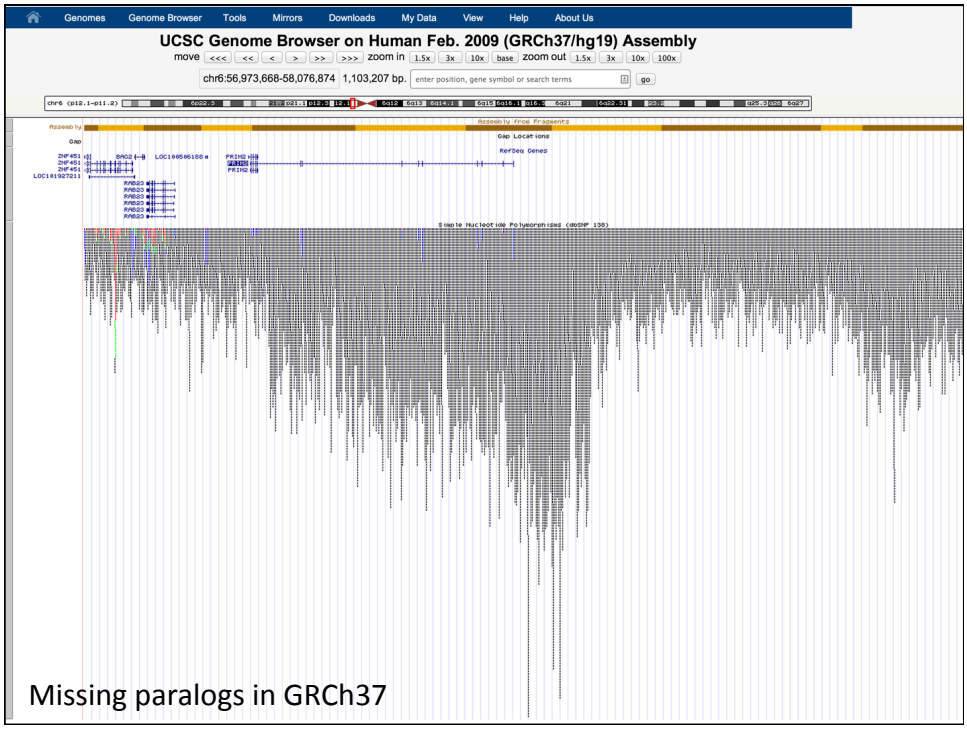


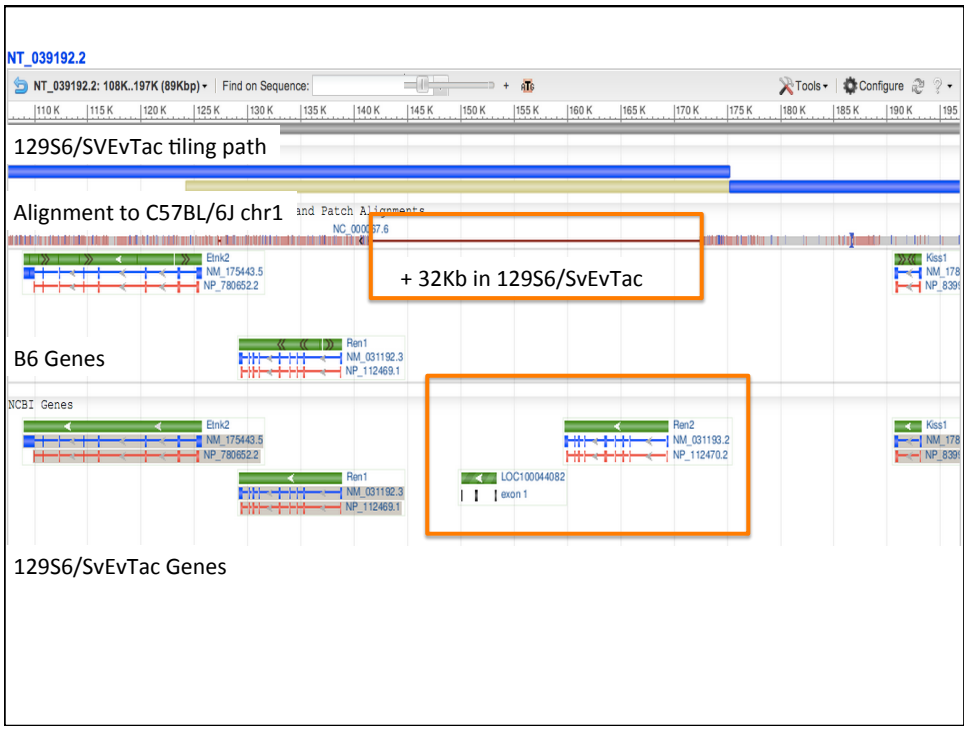
GRCh37

G>A (allelic difference – true variant)

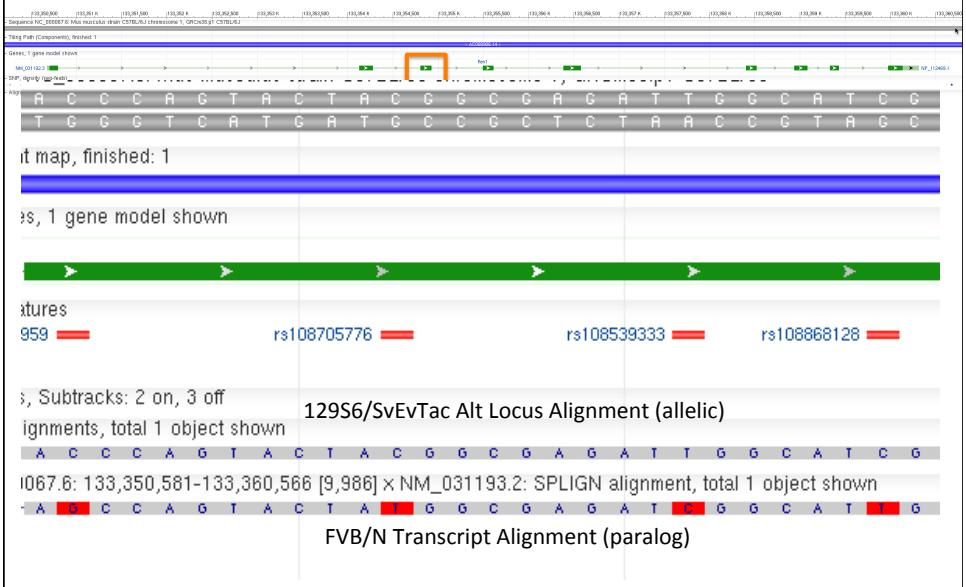
G>C (paralogous sequence variant- false positive)

May or may not detect increased coverage depending on sequencing depth
and library quality (easier to find with new technologies than with old, low through technologies)

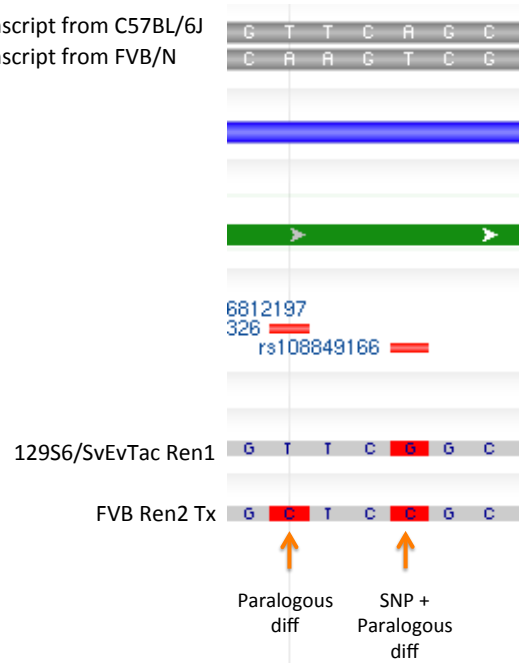

Genomes
Genome Browser
Tools
Mirrors
Downloads
My Data
View
Help
About Us
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
chr6:56,973,668-58,076,874 1,103,207 bp.
Missing paralogs in GRCh37

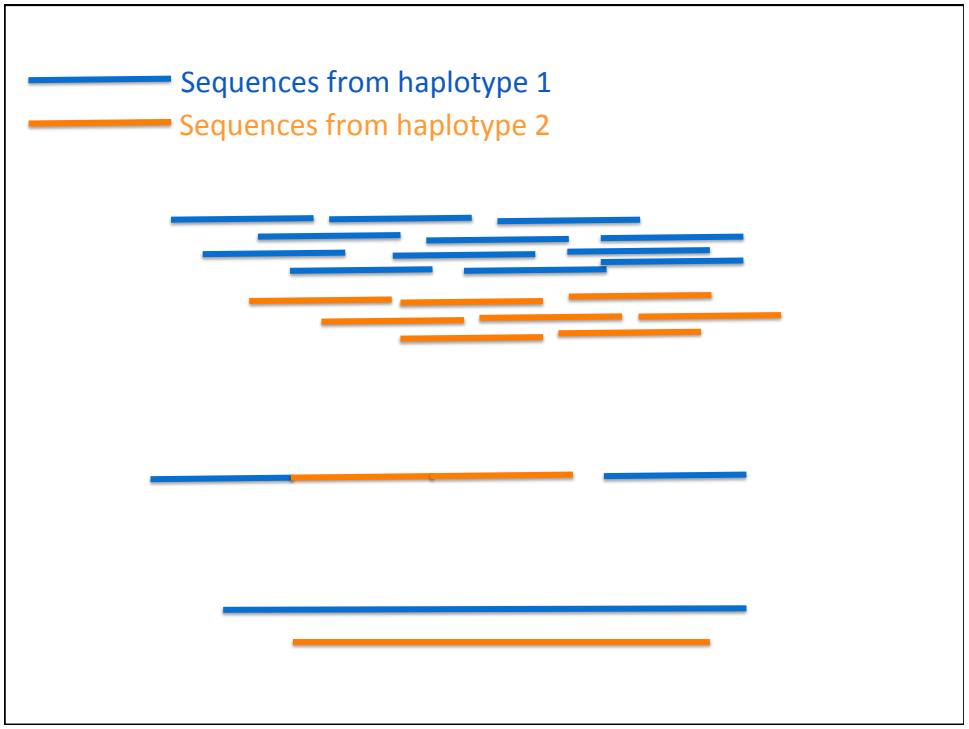

NT_039192.2
129S6/SVEvTac tiling path
Alignment to C57BL/6J chr1
+ 32Kb in 129S6/SvEvTac
B6 Genes
NCBI Genes
129S6/SvEvTac Genes

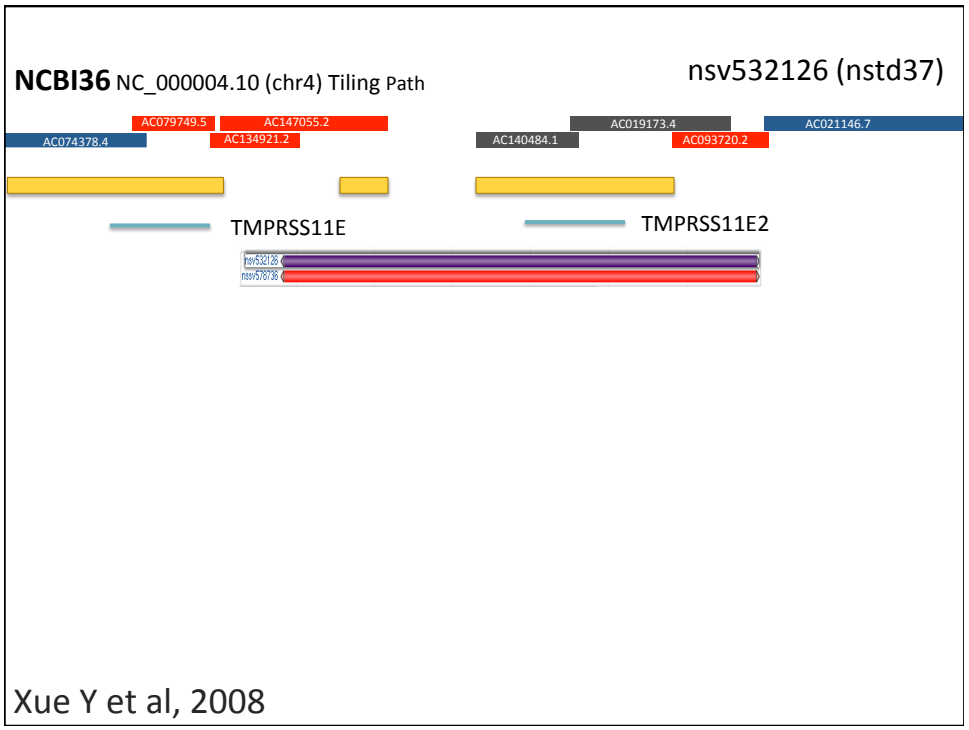
NM_031192.3: transcript from C57BL/6J
NM_031193.2: transcript from FVB/N



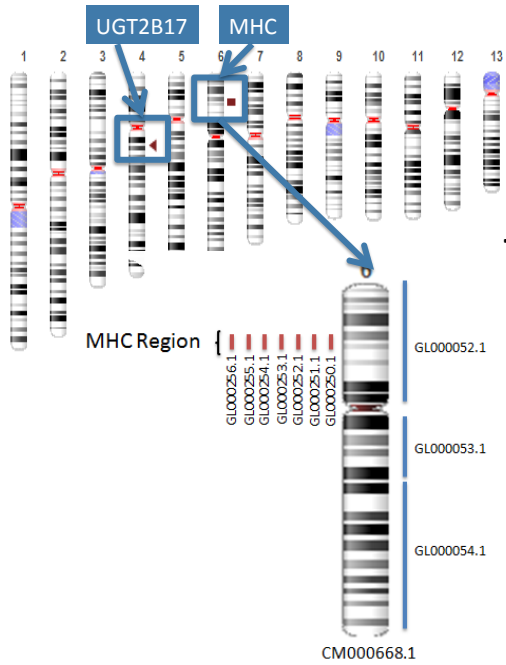
NM_031192.3: transcript from C57BL/6J
NM_031193.2: transcript from FVB/N







Xue Y et al, 2008

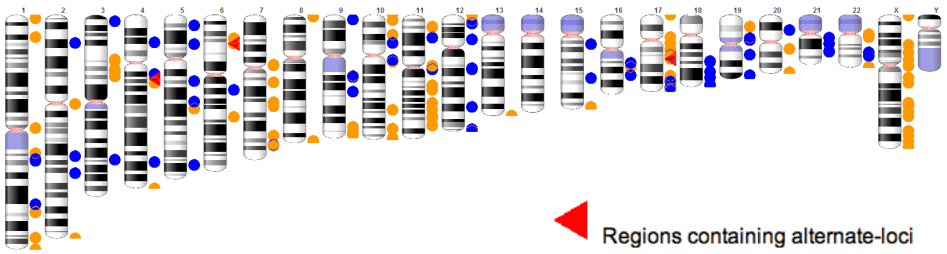


7 alternate haplotypes at the MHC

Alternate loci released as:
FASTA
AGP
Alignment to chromosome

http://genomereference.org

Updating the assembly



GRCh37.p13
(160 regions: >3% of chromosomes)

Regions containing alternate-loci
Regions containing fix patches
Regions containing novel patches

120 Fix PATCHES: Chromosome update in GRCh38

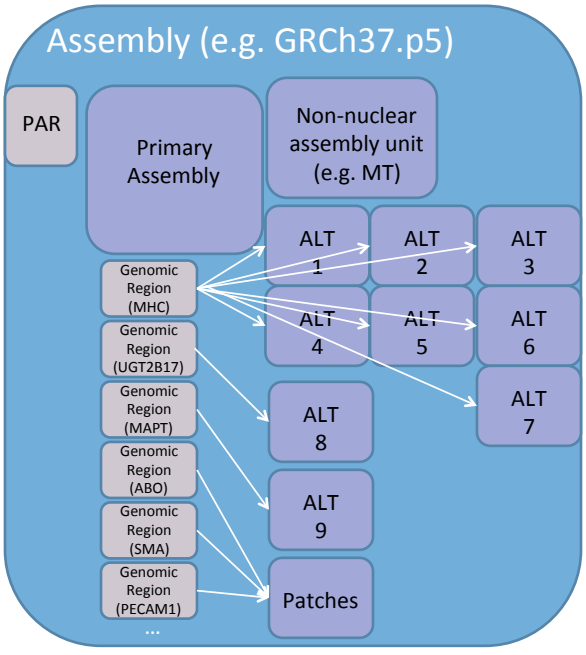
(adds >5 Mb of novel sequence to the assembly)

71 Novel PATCHES: Additional sequence added

(adds >800K of novel sequence to the assembly)

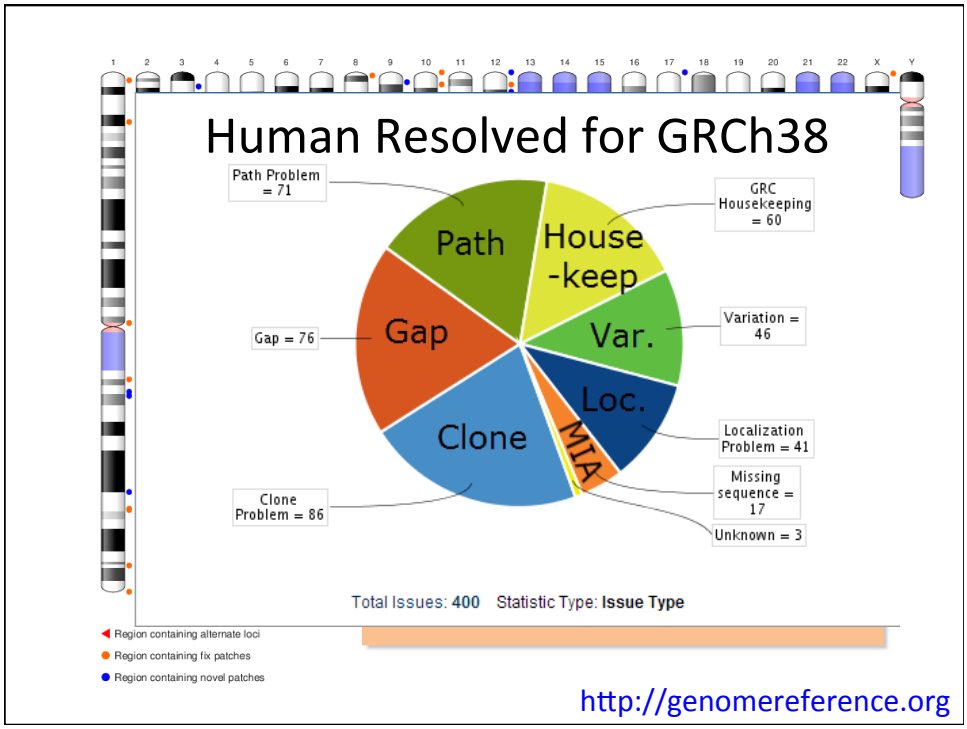
Releasing patches quarterly

Data Model

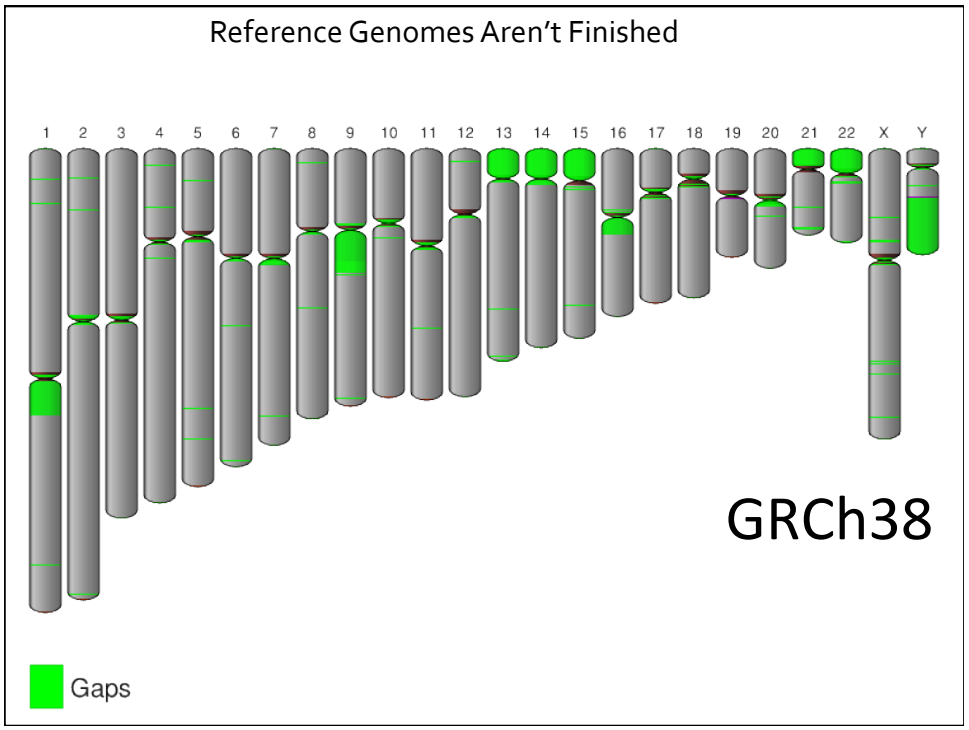


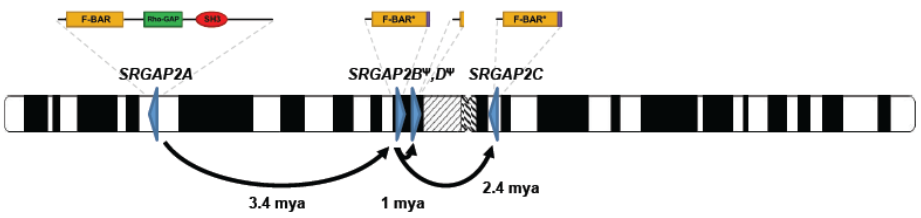
Data Model

Primary Assembly	GCA_000001305.1/ GCF_000001305.13
Non-nuclear assembly unit (e.g. MT)	GCA_000006015.1/ GCF_000006015.1
ALT 1	GCA_000001315.1/ GCF_000001315.1
ALT 2	GCA_000001325.1/ GCF_000001325.2
ALT 3	GCA_000001335.1/ GCF_000001335.1
ALT 4	GCA_000001345.1/ GCF_000001345.1
ALT 5	GCA_000001355.1/ GCF_000001355.1
ALT 6	GCA_000001365.1/ GCF_000001365.2
ALT 7	GCA_000001375.1/ GCF_000001375.1
ALT 8	GCA_000001385.1/ GCF_000001385.1
ALT 9	GCA_000001395.1/ GCF_000001395.1
Patches	GCA_000005045.5 GCF_000005045.4



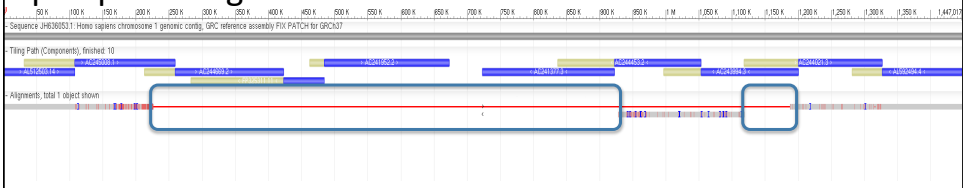

GRCh38 is here!
(Dec, 2013)
HBO Game of Thrones


Reference Genomes Aren't Finished
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
GRCh38
Gaps



1q32 1q21 1p21

1p21 patch alignment to chromosome 1

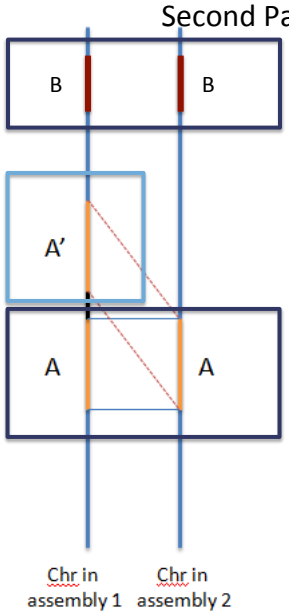


Dennis et al., 2012

Migrating to GRCh38

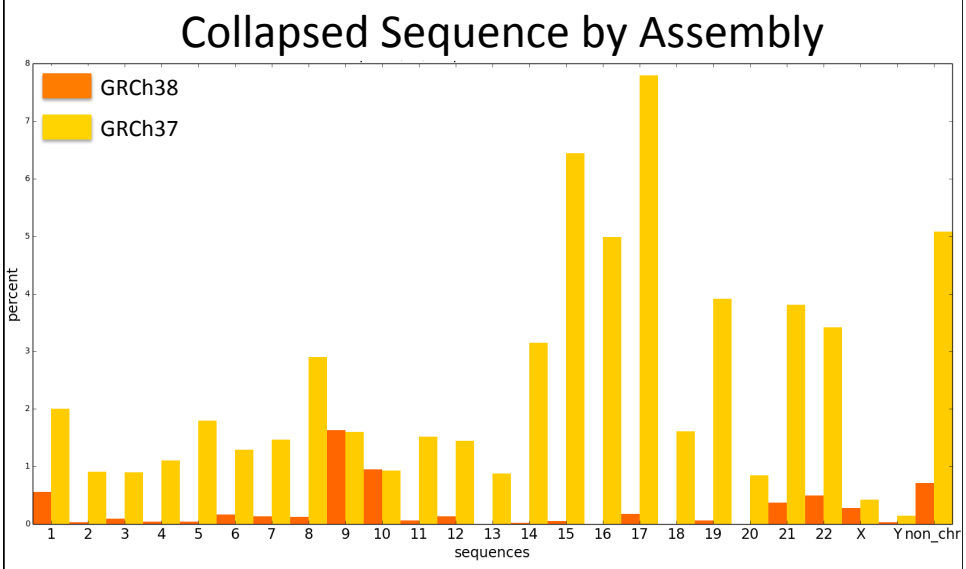
First Pass + Second Pass = Collapse
Second Pass Only = Expansion

NCBI Remap tool

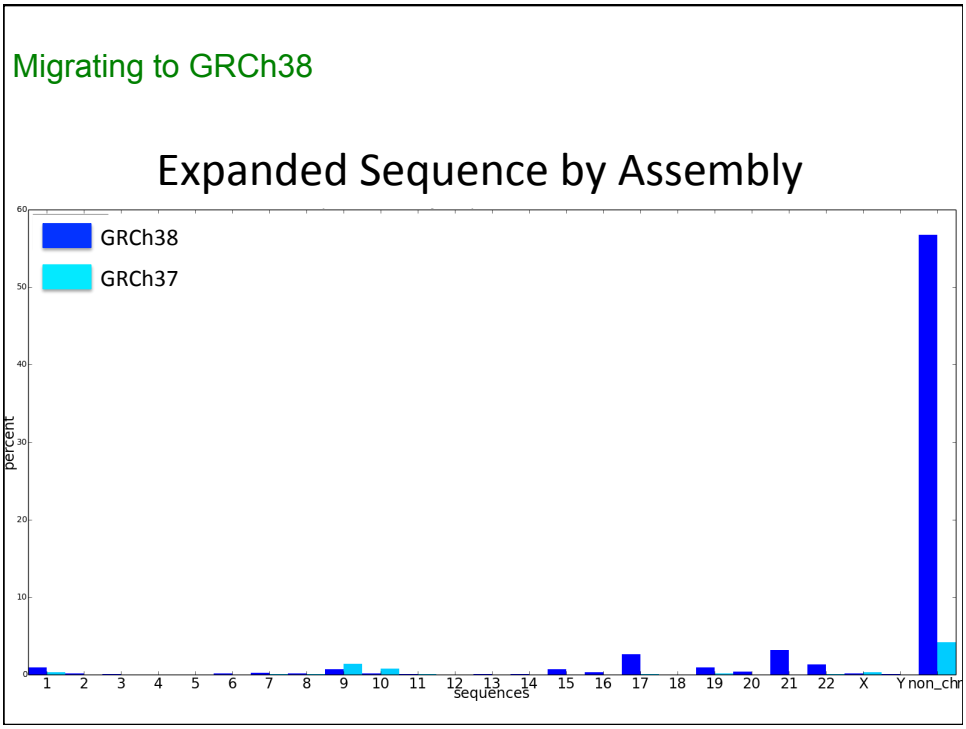


http://www.ncbi.nlm.nih.gov/genome/tools/remap

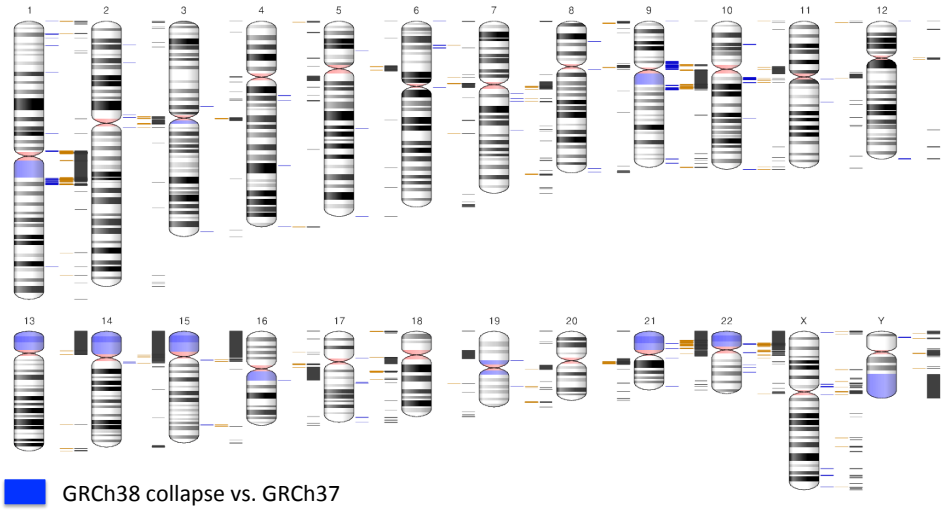
Migrating to GRCh38



Migrating to GRCh38



Migrating to GRCh38



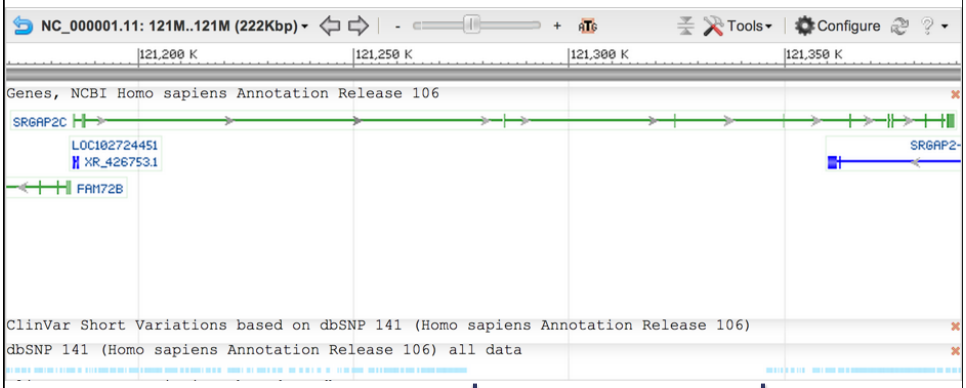
GRCh38 collapse vs. GRCh37

GRCh38 expansion vs. GRCh37

GRCh38 no alignment vs. GRCh37

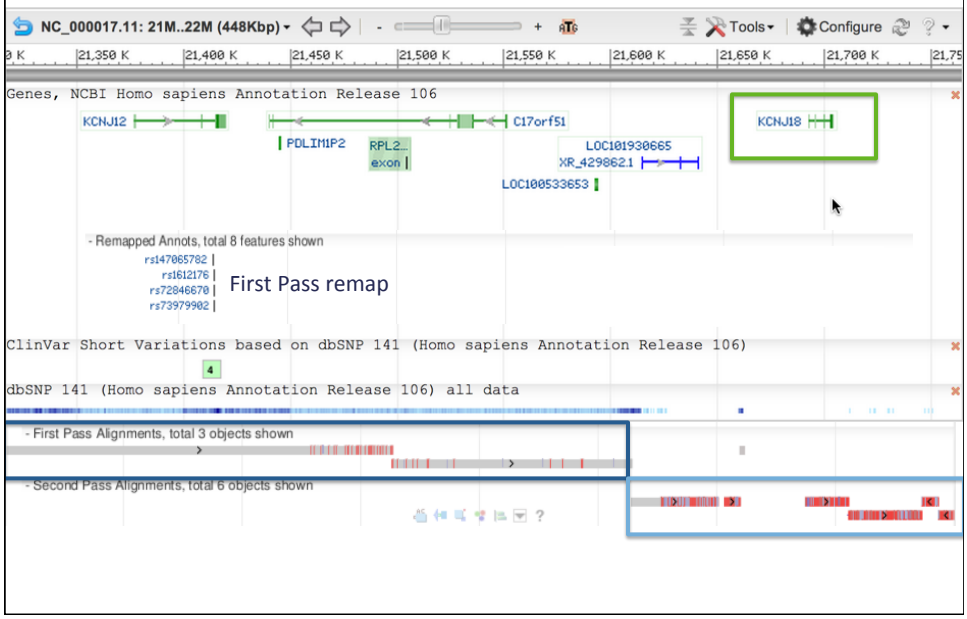
Migrating to GRCh38

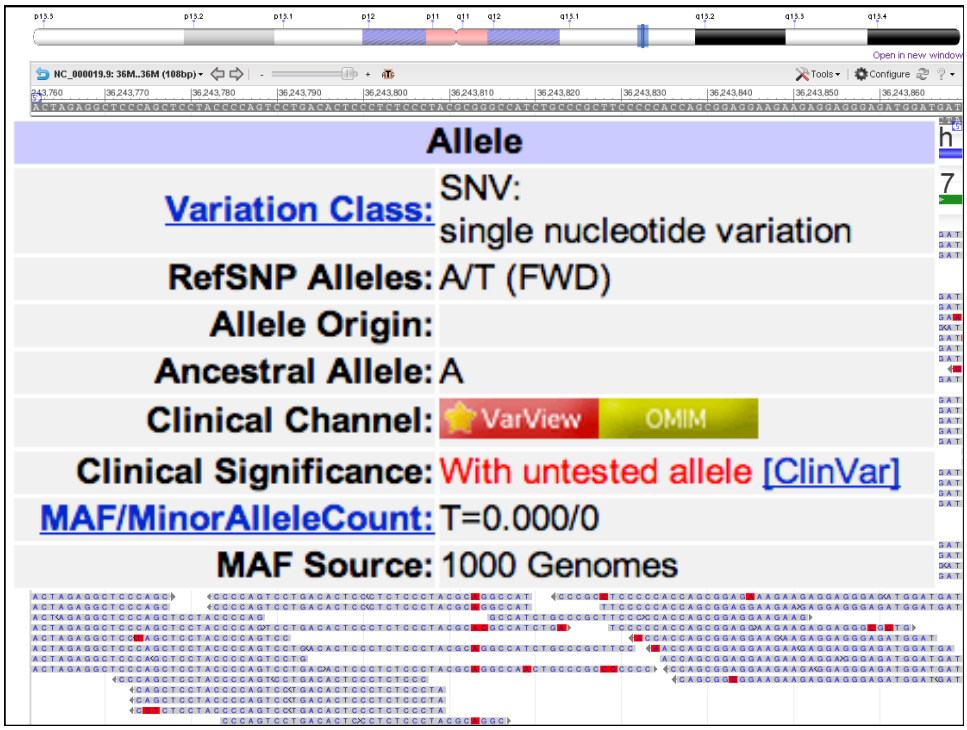
GRCh38 annotation (NCBI)



No dbSNP 141 variants

Migrating to GRCh38

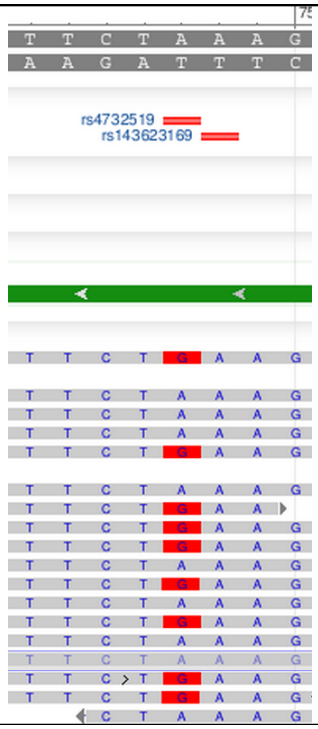




A = 0.000
G=1.000

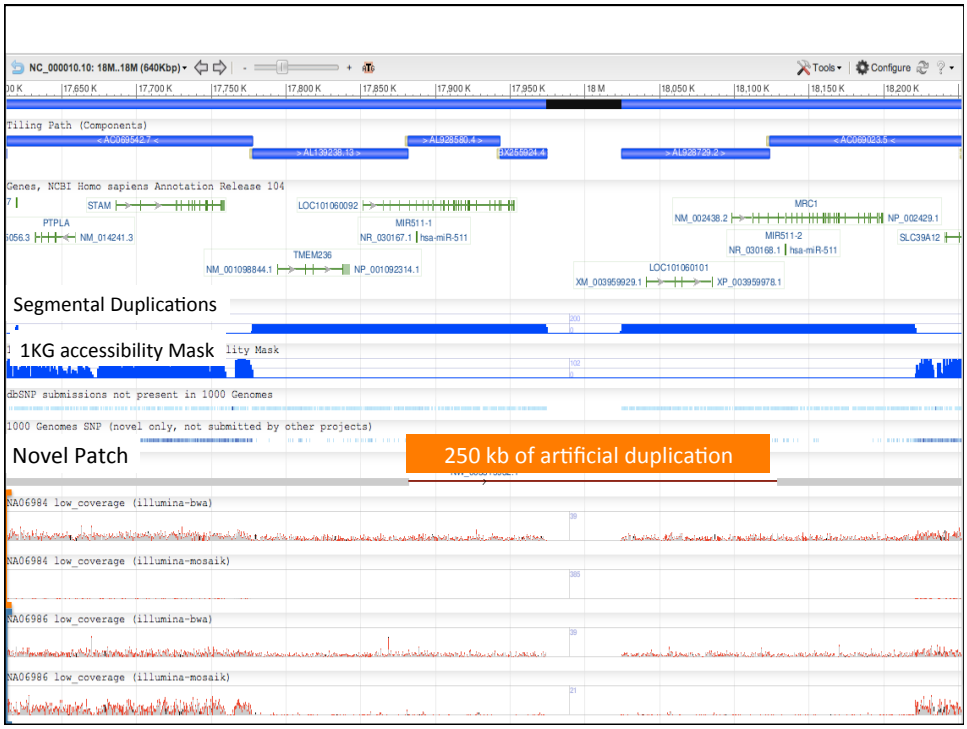
rs4732519

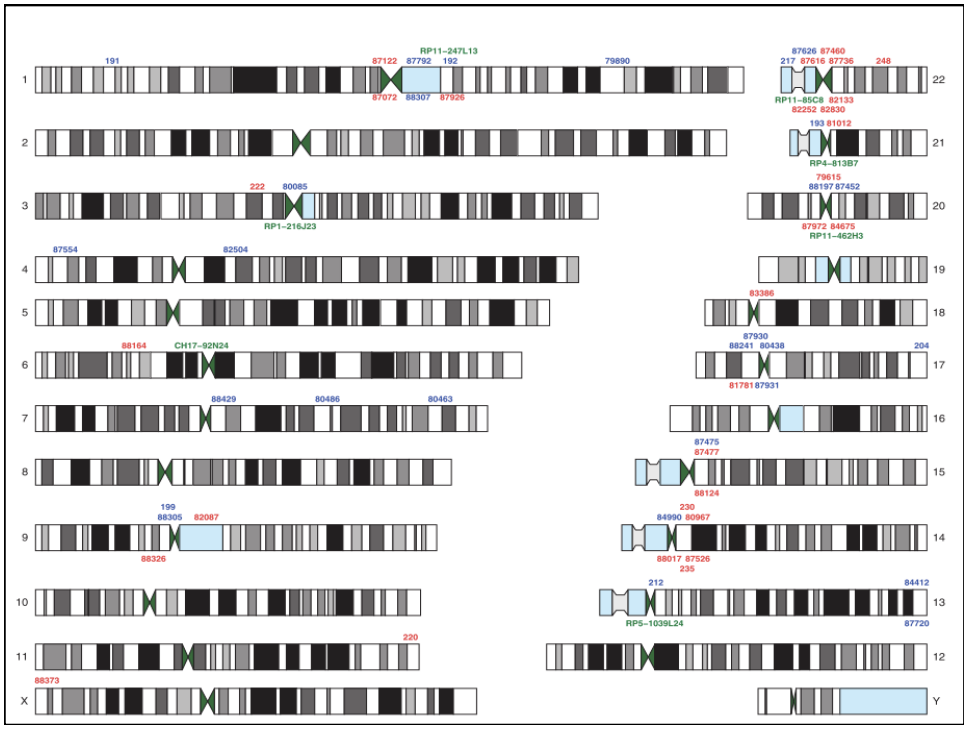
RP11 WGS reads

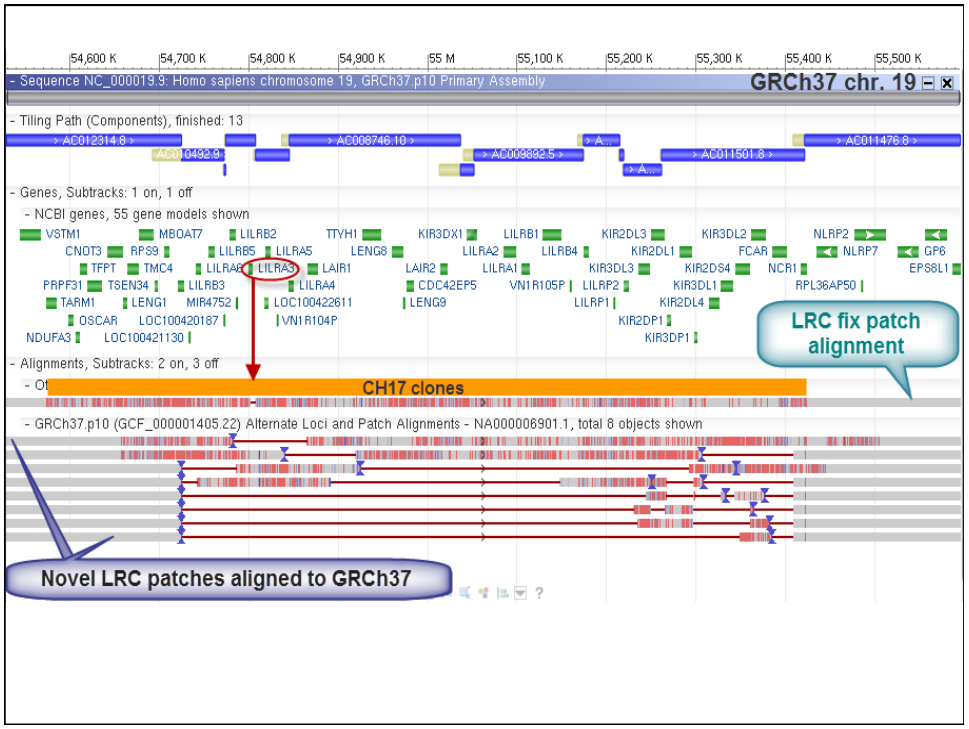


Private RP11 variant?
Missing in 1000G?

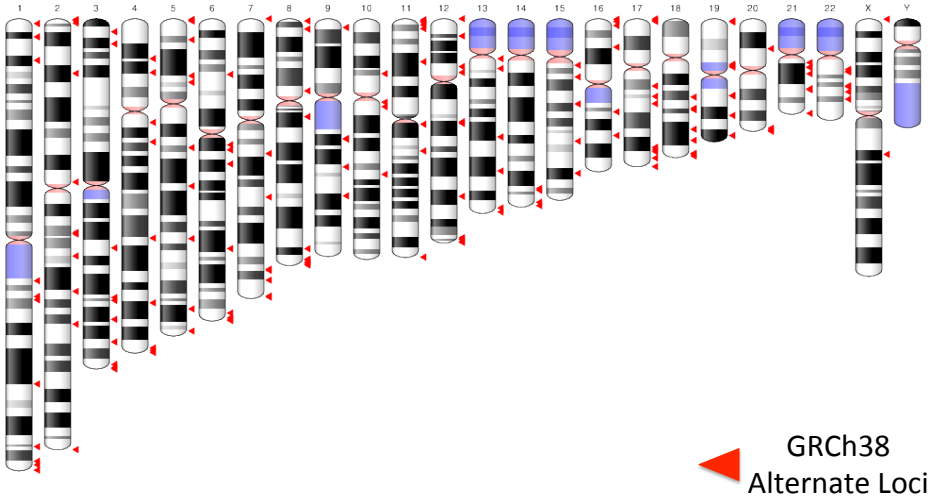
rs4732519

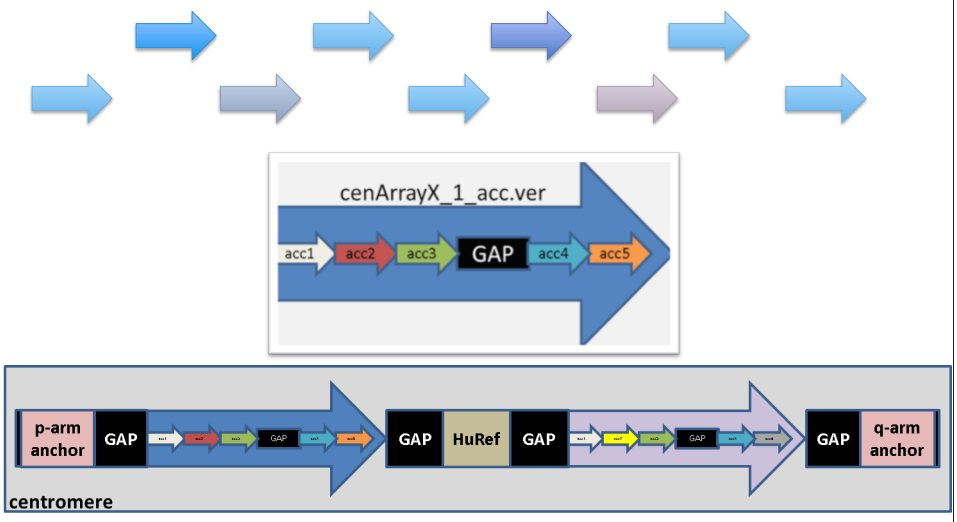

Segmental Duplications
1KG accessibility Mask
Novel Patch
250 kb of artificial duplication





Migrating to GRCh38





Karen Hayden and Jim Kent

Take home messages

- The human and mouse reference assemblies are not linear single haplotype representations
- Complex variation and segmental duplication makes genome assembly challenging.
  - This is true even using newer long read data
- New assembly will challenge our understanding of annotation