Logistics

Next two weeks class will be held via Zoom

Quiz 1 next Monday, 9/19/22 at beginning of class

- We will meet via Zoom at regular time then I will launch the quiz
- Multiple choice via Brightspace
- Will cover materials presented in class through today
- Quiz will be available for only 20-30 minutes, no late submissions.

Example quiz question available today by tomorrow

Homework 1 will be released by next Wednesday

SNP-calling vs. genotyping

“SNP”-calling: establish whether a nucleotide position is polymorphic (variable) in a population

Variant-calling: SNP-calling + calling of other types of variants (e.g., indels, structural variants)

Genotyping: Calling the genotype of an individual sample

Genotyping and SNP-calling are often used interchangeably

reference genome			diploid samples 1	2	3	4
chr1	10,002	G	G/G	G/T	G/T	T/T
chr1	11,112	A	A/A	A/G	A/A	A/A
chr1	11,137	A	-/-	A/-	-/-	A/A

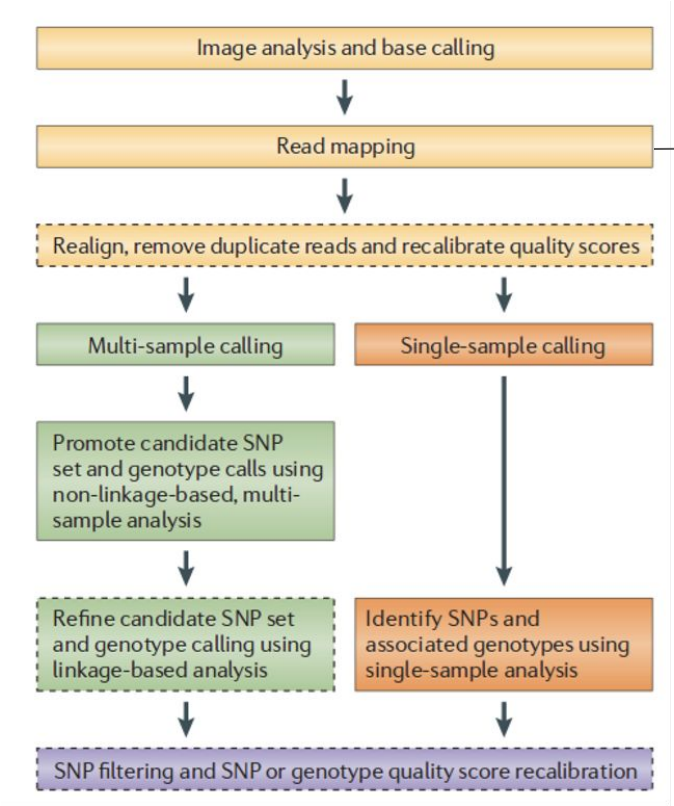
genotype of sample 4

(“T” is “alternate”. or non-reference, base for this SNP)

SNPs at positions 10,002 and 11,112

“indel” (=insertion/deletion) polymorphism at position 11,137

“re-sequencing” approach with Illumina short read data



Nielsen et al., Nature Reviews Genetics (2011)

Genotyping technologies

Historically many technologies used in human population genetics

Genotyping array (“SNP-Chip”) is commong

Discover polymorphic sites by sequencing a reference panel

Build SNP-Chip based on sites discovered in reference panel

This strategy introduces an ascertainment bias

The human reference genome

A haploid representation of the human genome

A consensus DNA sequence derived from 17 individuals from Buffalo

Consists of complete or near-complete chromosome sequences (“pseudomolecules”)

The reference genome in FASTA format

FASTA formatted sequence represents plus strand only (by convention)

Example:

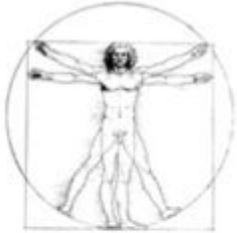
5’-ATGCGGGGGCCCAATA-3’ (plus)

3’-TACGCCCCCGGGTTAT-5’ (minus)

FASTA representation:

```
>chromosome_id
ATGCGGGGGCCCAATA
```

GRCh38 (=hg38)



Human

The human genome assembly was produced as part of the Human Genome Project (HGP). The previous assembly (NCBI36) was the last one produced by the HGP and was described in 2004 (PMID: 15496913); this was the starting point for the GRC. The assembly is based largely on assembling overlapping clone sequences.

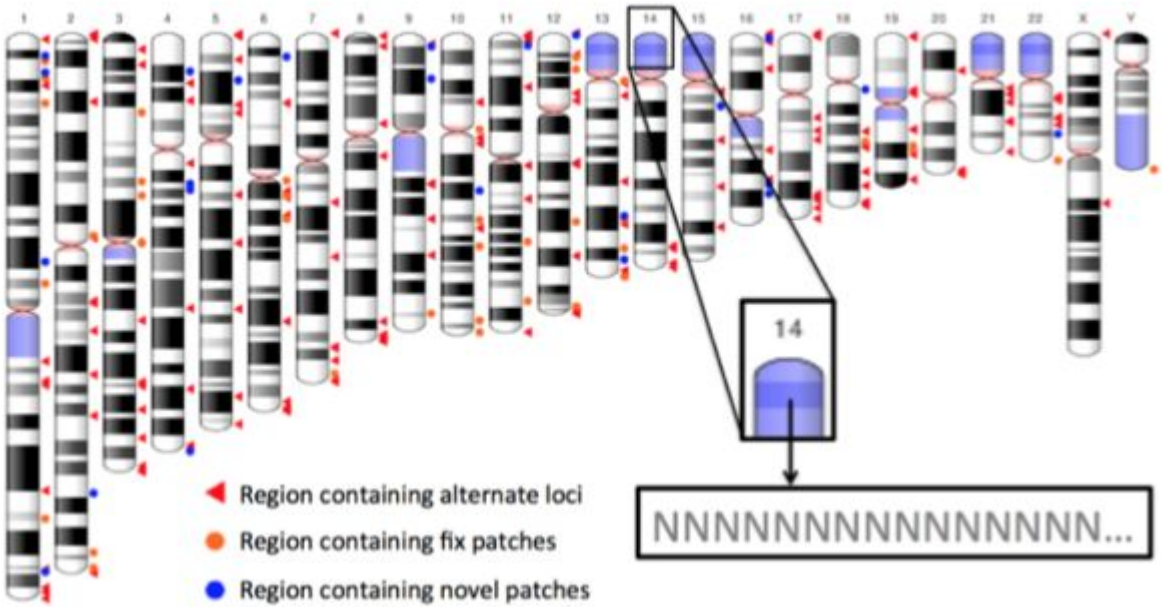
Human assembly information

Current major assembly	GRCh38
Regions with alternate loci	178
Assembly N50	67,794,873 bp
Remaining gaps	875
Patch release version	p11
Patches released	FIX: 64, NOVEL: 59

https://www.ncbi.nlm.nih.gov/grc/human

https://software.broadinstitute.org/gatk/documentation/article?id=7857

GRCh38



https://www.ncbi.nlm.nih.gov/grc/human

https://software.broadinstitute.org/gatk/documentation/article?id=7857

GRCh38/hg38 components

Unlocalized sequence: associated with a chromosome but unknown location and orientation

Unplaced sequence: not associated with a chromosome

Decoy sequences: sequences that are are included as a “sink” in the FASTA-formatted file for read alignment.

Example: Epstein-Barr Virus (EBV)

Alternate sequences: sequence variants from hypervariable parts of the genome (e.g., MHC)

Masked regions (either “hard” or “soft”)

Examples: transposable elements or pseudoautosomal regions (PAR)

https://www.ncbi.nlm.nih.gov/grc/human

https://software.broadinstitute.org/gatk/documentation/article?id=7857

Patches to GRCh38/hg38

Reference stability is important

Periodically GRC releases minor versions or “patches”

Patches new scaffold sequences that either fix errors or extend the original sequence into gaps

Importantly, the chromosome sequences are not modified at the time of patch release

Rather, the new scaffolds are accessioned and made available as raw sequences and alignments to existing chromosomes

Genome Reference Consortium

GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Overview | Definitions | FAQ | Patches | Human Region Examples | Mouse Region Examples | Workshops

Introduction to Patches

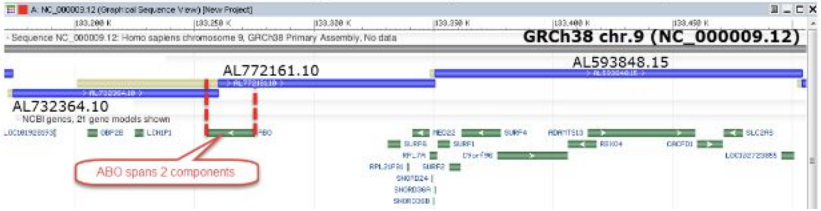
- What are patches?
- What types of patches are there?
- Do patches result in changes to chromosome coordinates?
- What is a patch release?
- How often does the GRC release patches?
- Why does the GRC release patches?
- How can I tell if an assembly update is a patch release or a major release?
- How should I refer to patches in a publication?
- Where can I find the list of assembly regions that have been patched?
- What implications do patches have for my analyses?

What are patches?

Patches are accessioned scaffold sequences that represent assembly updates. They add information to the assembly without disrupting the chromosome coordinates. Patches are given chromosome context via alignment to the current assembly. Together, the scaffold sequence and alignment define the patch. Patch sequences and alignments can be downloaded from the GenBank FTP site.

What types of patches are there?

- **FIX patches:** Fix patches represent changes to existing assembly sequences. These are generally error corrections (addressed by approaches such as base changes, component replacements/updates, switch point updates or tiling path changes) or assembly improvements, such as the extension of sequence into gaps. A fix patch scaffold represents a preview of what the assembly will look like at the next major (coordinate changing) release. When the next major release occurs, the accessions for the fix patch scaffolds will be deprecated and the changes will be found in the chromosomes. An example of a fix patch is shown in Figure 1.



https://www.ncbi.nlm.nih.gov/grc/help/patches/#coords

How to choose a GRCh38/hg38 reference genome?

Example:

Ensembl has 900+ fasta files representing GRCh38 (!)

Files vary in masking and presence/absence of alternate sequences or consist of single chromosomes

How to choose a GRCh38/hg38 reference genome?

genomespot.blogspot.com/2015/06/mapping-ngs-data-which-genome-version.html

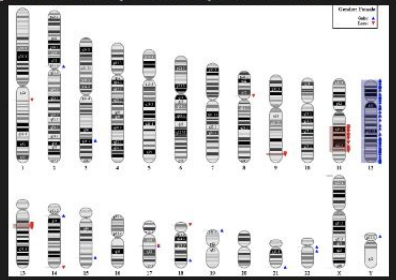
Genome Spot

Practical tips for genome analysis

Mapping NGS data: which genome version to use?

June 02, 2015

Aligning reads to the genome is a key step in nearly all NGS data pipelines, the quality of an alignment will dictate the quality of the final results. So for beginners in this space, the options available can be a bit overwhelming.



Which options are available?

Depending on what species you are working on, you will have either a limited number of choices or a vast number of choices. These include NCBI, Ensembl, UCSC as well as the consortia that generate these genome builds, such as the

Some sources of reference genome:
Ensembl
UCSC (“analysis ready” fasta)
Encode project

Current version GRCh38/hg38

Old builds

GRCh37/b37 and Hg19
GRCh36/b36 and Hg18

Important to always to consider the exact build used in upstream analysis

Excellent description of the legacy builds at the GATK (Broad) website:

https://gatk.broadinstitute.org/hc/en-us/articles/360035890951-Human-genome-reference-builds-GRCh38-or-hg38-b37-hg19

SNP “rs numbers”

Reference SNP (“rs” = “refSNP”) identifiers in dbSNP database (NCBI)

650,000 rs numbers have been assigned to specific variants

Linked to ClinVar (clinically significant variant database)

Example: Search dbSNP for “rs328”

1. Go to https://www.ncbi.nlm.nih.gov/snp/
2. Enter rs identifier
3. Click on rs328 search result

dbSNP Short Genetic Variations

Search for terms | Search
Examples: rs268, BRCA1 and more | Advanced search

Welcome to the Reference SNP (rs) Report
All alleles are reported in the Forward orientation. Click on the Variant Details tab for details on Genomic Placement, Gene, and Amino Acid changes. HGVS names are in the HGVS tab.

Reference SNP (rs) Report

Switch to classic site

Download

rs328

Current Build 155
Released April 9, 2021

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr8:19962213 (GRCh38.p13)	**Gene : Consequence**	LPL : Stop Gained
Alleles	C>A / C>G	**Publications**	121 citations LitVar 371
Variation Type	SNV Single Nucleotide Variation		
Frequency	G=0.087925 (23273/264690, TOPMED) G=0.092156 (23148/251182, GnomAD_exome) G=0.089338 (12522/140164, GnomAD) (+ 23 more)	**Genomic View**	See rs on genome

FEEDBACK

- **Variant Details**
- Clinical Significance
- Frequency
- HGVS
- Submissions
- History
- Publications
- Flanks

Genomic Placements

Sequence name	Change
GRCh37.p13 chr 8	NC_000008.10:g.19819724C>A
GRCh37.p13 chr 8	NC_000008.10:g.19819724C>G
GRCh38.p13 chr 8	NC_000008.11:g.19962213C>A
GRCh38.p13 chr 8	NC_000008.11:g.19962213C>G
LPL RefSeqGene (LRG_1298)	NG_008855.2:g.65497C>A
LPL RefSeqGene (LRG_1298)	NG_008855.2:g.65497C>G

Gene: LPL, lipoprotein lipase (plus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
lipoprotein lipase precursor	NP_000228.1:p.Ser474Ter	S (Ser) > * (Ter)	Stop Gained
lipoprotein lipase precursor	NP_000228.1:p.Ser474Ter	S (Ser) > * (Ter)	Stop Gained
LPL transcript	NM_000237.3:c.1421C>A	S [TCA] > * [TAA]	Coding Sequence Variant
LPL transcript	NM_000237.3:c.1421C>G	S [TCA] > * [TGA]	Coding Sequence Variant

Variant Call Format (VCF)

VCF is a rich format that accomodates all forms of sequence variation (SNPs, indels, structural variants of any type)

Universal standard in NGS-based analysis of sequence variation

Detailed specification suitable for automated processing and computation

Variant Call Format (VCF)

Example: VCF from v4.2 specification

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

from: Variant Call Format Specification (https://samtools.github.io/hts-specs/VCFv4.2.pdf)

Variant Call Format (VCF)

Example: Filters are applied to identify low quality SNPs

```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

from: Variant Call Format Specification (https://samtools.github.io/hts-specs/VCFv4.2.pdf)

Variant Call Format (VCF)

Example: Column dependencies

INFO	FORMAT	NA00001	NA00002	NA00003
NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

sum of sample read depths 1 + 8 + 5 = 14 in INFO column DP field

Column-wise dependencies exist, such that if you remove a column, many INFO column tags must be updated

GATK has tools for performing such operations

from: Variant Call Format Specification (https://samtools.github.io/hts-specs/VCFv4.2.pdf)

Parsing VCF

Example: VariantAnnotation package (R/Bioconductor)