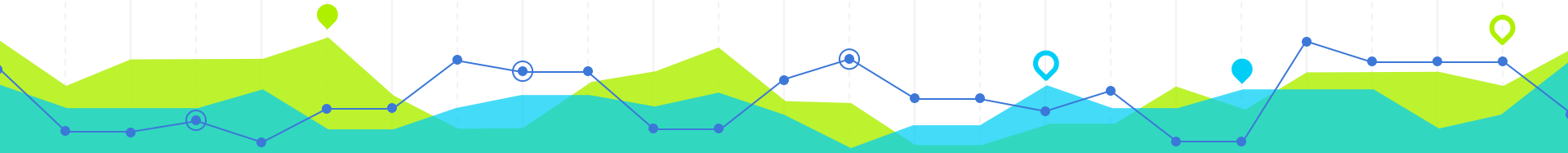
Barriers to Reference Genome Assembly Migration for Clinical Genetics Laboratories



Reference genome assembly

1

Genome Reference Consortium

- HGP > GRC 2007
- www.ncbi.nlm.nih.gov/grc/human

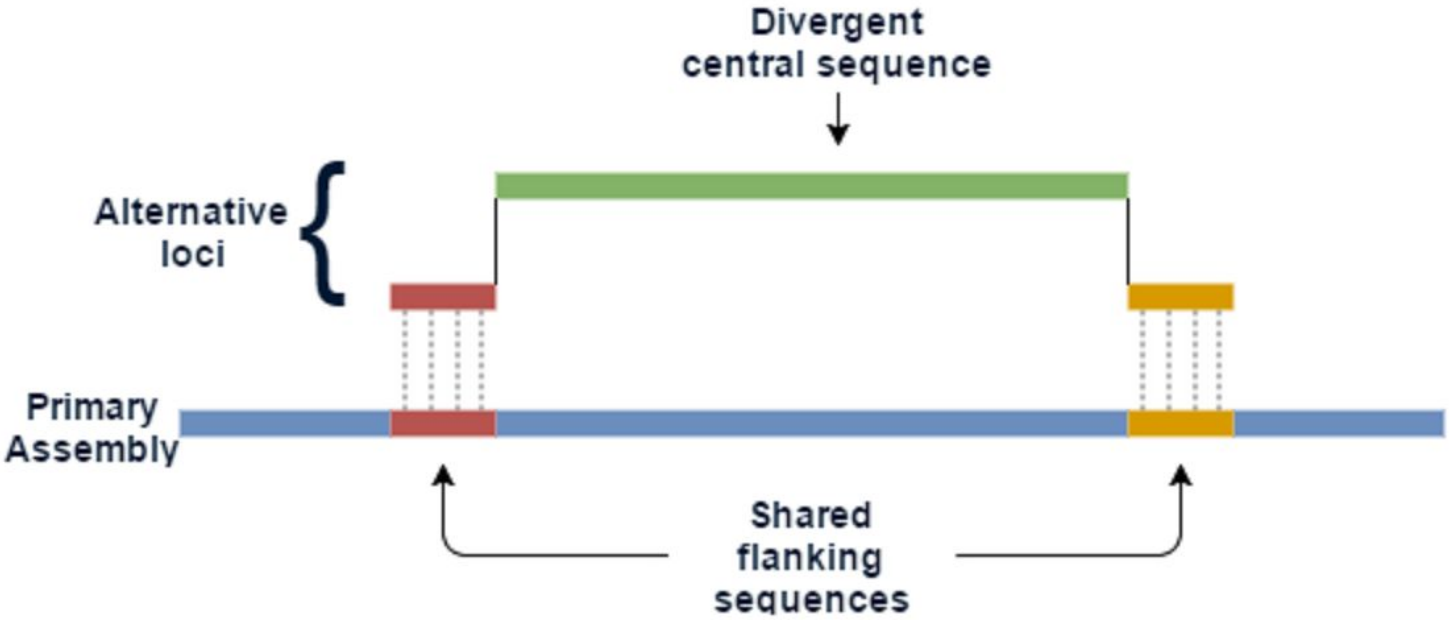
- Major releases every few years
  - GRCh37 2009
  - GRCh38 2013
- Minor releases more frequently
  - GRCh38.p13

GRC Reference Assembly Model

- **Primary**
  - Chromosomes
  - Unlocalised sequences
  - Unplaced sequences
- **Alternate loci**
- **Patches**
  - Fix
  - Novel

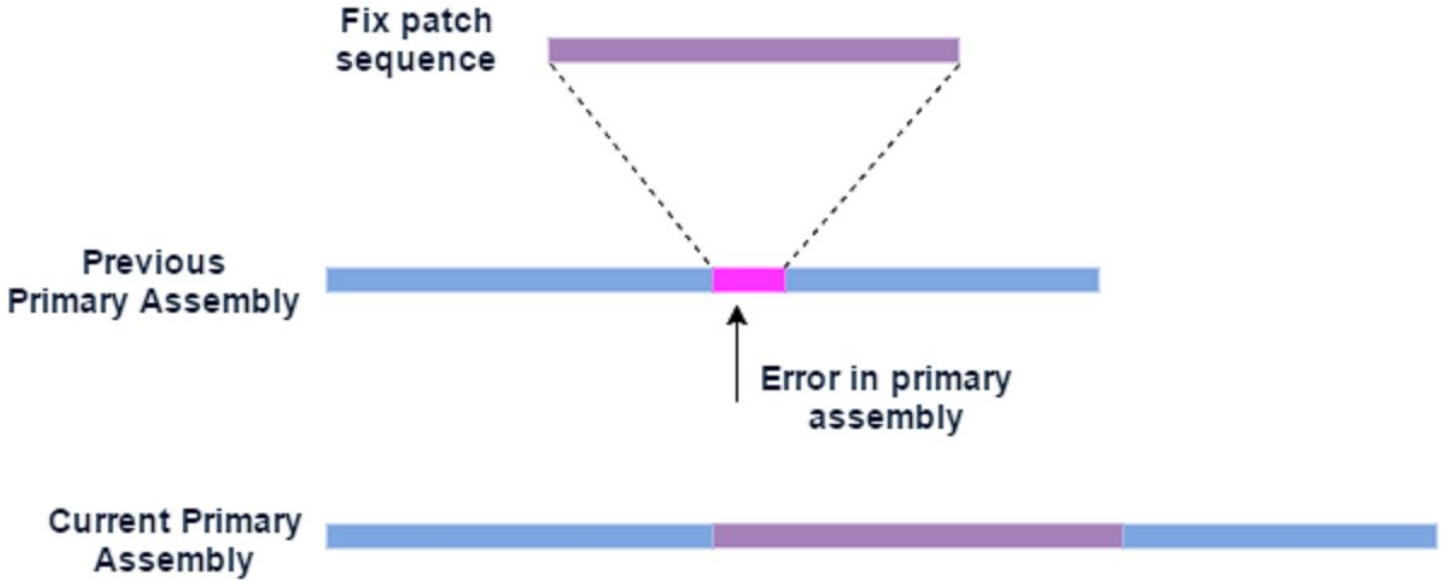
Reference assembly model
Alternate loci

- Regions which differ significantly from the primary within some individuals
- Enables representation of complex structural diversity within the population



Reference assembly model
Patch sequences

- Novel patches - additional alternate loci
- Fix patches - correct errors



Applications of the reference genome in clinical genomics

Assembly/alignment

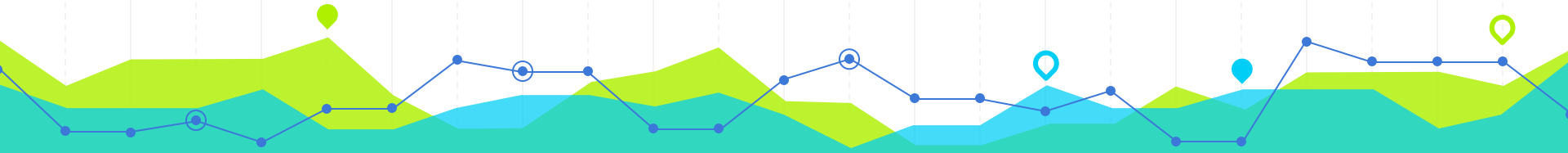
Individual patient genomes are reassembled using the reference genome as a template

Variant calling

Variants are identified as differences between the assembled genome and the reference genome

Annotation/interpretation

Stable coordinate system provided by the reference enables existing knowledge to be assigned consistently to a specific region of the genome



GRCh37 issues 2

GRCh37 issues
Mosaic Haplotypes

Description

Mosaic haplotypes within the reference assembly which are highly unlikely to be seen in nature

Cause

BAC contigs in the assembly tiling path are derived from different individuals or chromosomes

GRCh38 fix

Regions identified and resolved using haploid genome assembly data from hydatidiform mole

GRCh37 issues
Gaps

Description

Gaps ('N' nucleotides) constitute 239Mb, or 7.6% of the GRCh37 assembly.

Cause

Regions of the genome which are recalcitrant to sequencing and/or assembly, often due to their repetitive nature

GRCh38 fix

Gaps reduced to 160Mb, or 5.0%, by addition of 78Mb of additional sequence with no GRCh37 alignment.

Centromeres represented by model sequences.

GRCh37 issues
Complex Structural Variation

Description

The diversity of structural variation present within the population in highly variable regions of the genome is not represented within the reference assembly

Cause

The linear structure of the primary assembly can only represent a single haplotype at a given genomic location.

GRCh38 fix

Expanded the number of alternative loci scaffolds from 9 scaffolds within 3 regions in GRCh37, to 261 scaffolds within 178 regions in GRCh38.

GRCh37 issues
Minor alleles

Description

Alleles with low population allele frequency, including pathogenic alleles, are present within the reference assembly

Cause

Individual genomes used to construct the reference assembly contain minor alleles and miscalled bases which are incorporated into the assembly

GRCh38 fix

6182 SNVs
910 deletions
489 insertions

present within GRCh37 but absent from 1000 Genomes data have been corrected in GRCh38

GRCh37 issues
Misassembly errors

Description

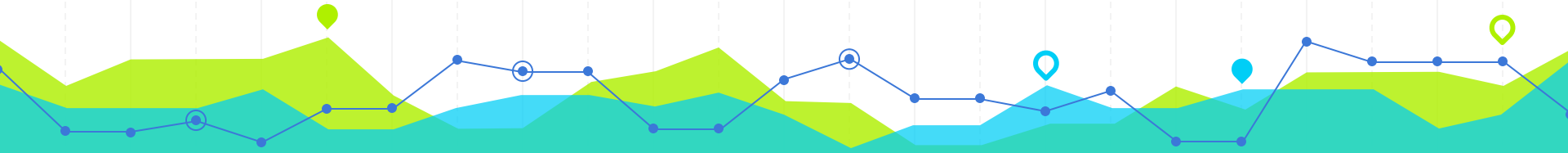
Misassembled regions of the genome

Cause

Some regions of the genome are inherently difficult to reassemble, often due to their repetitive nature, resulting in errors during genome assembly

GRCh38 fix

Misassembled regions such as 1q21 and 10q11 have been retiled.



Migration barriers

3

What are the barriers preventing migration to GRCh38?

- GRCh38 addresses many known issues with GRCh37
- Variant detection is improved with GRCh38
- GRCh38 released in 2013, yet is not being used by NHS labs

- What are the barriers preventing migration?
  - Census survey
  - Semi-structured interviews
  - Performed in 2015/16

Census survey

Practical issues

- Software
- Testing strategies
- Lab throughput
- Resources:
  - Staffing
  - Computational

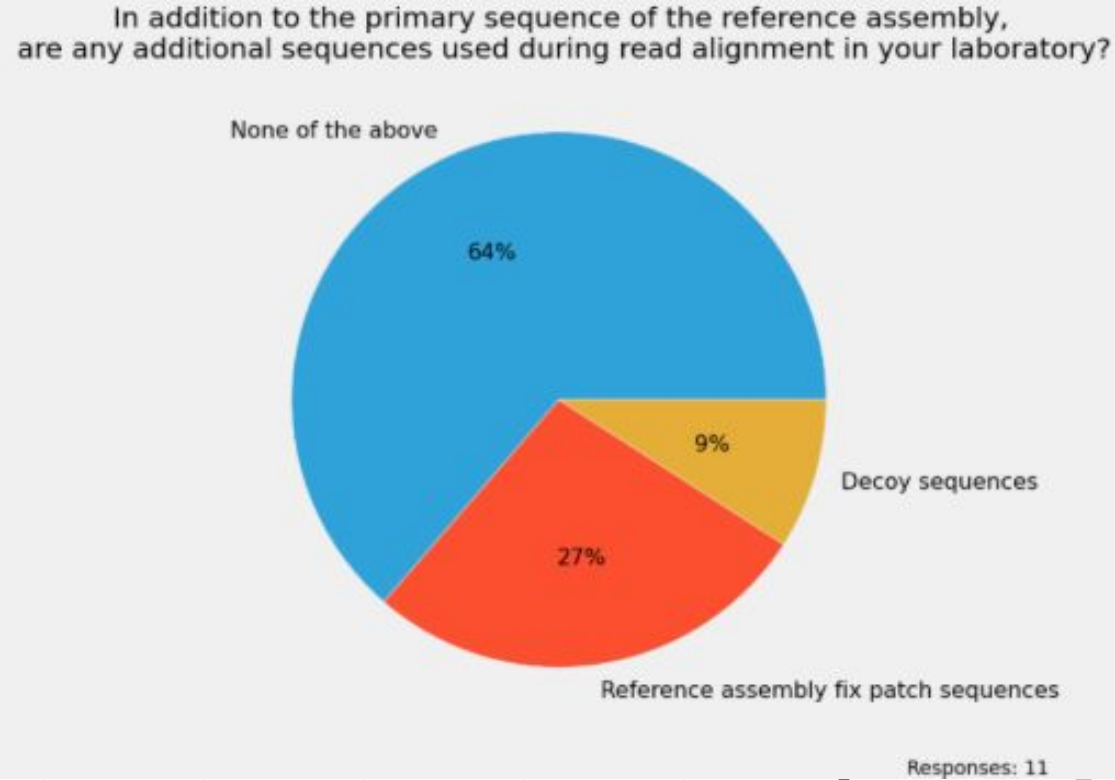
Opinions/perceptions

- Understanding
- Impact on results
- Feasibility
- Management of existing data

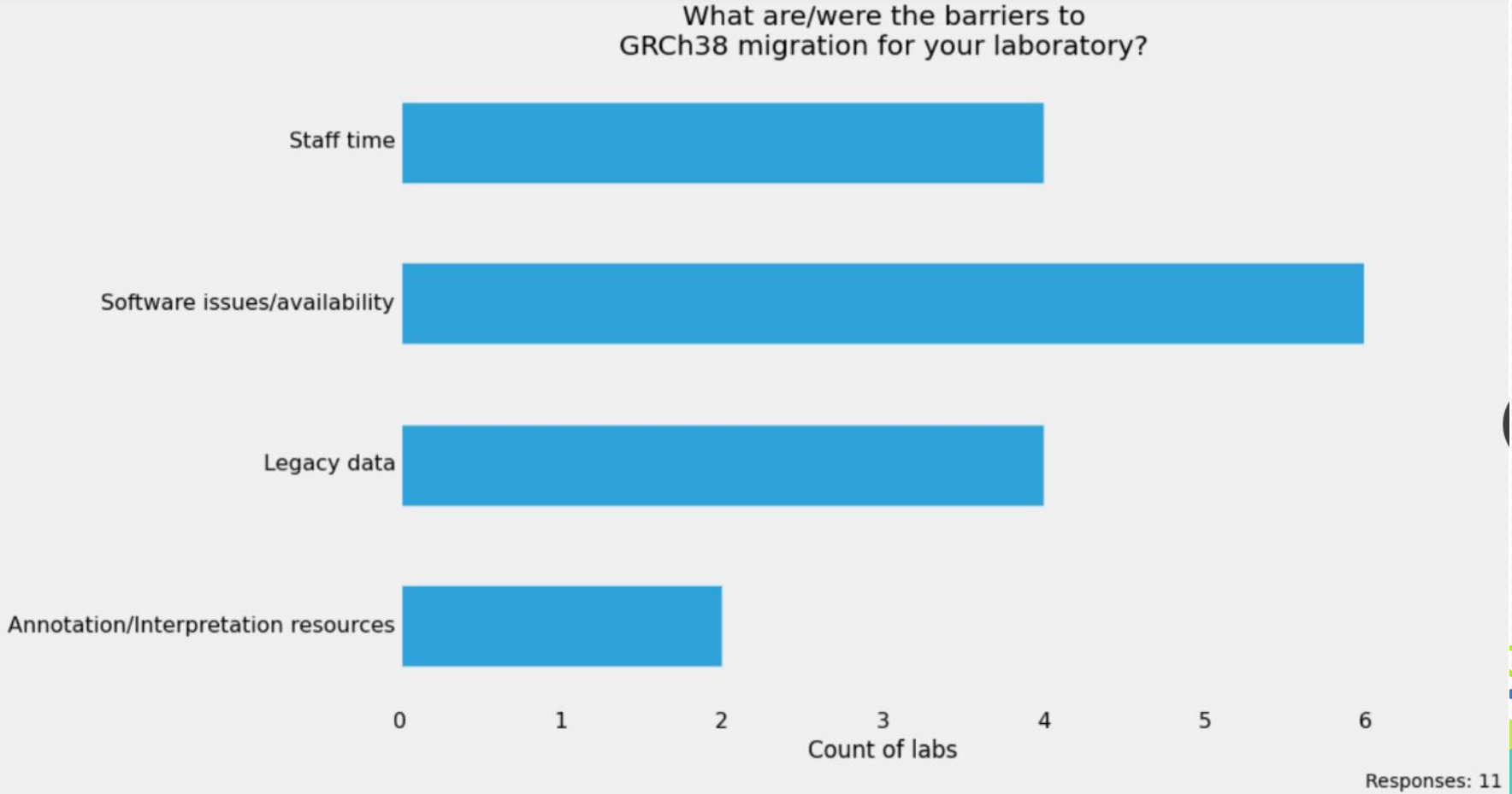
Census survey findings 1

- All using GRCh37
- None using GRCh38

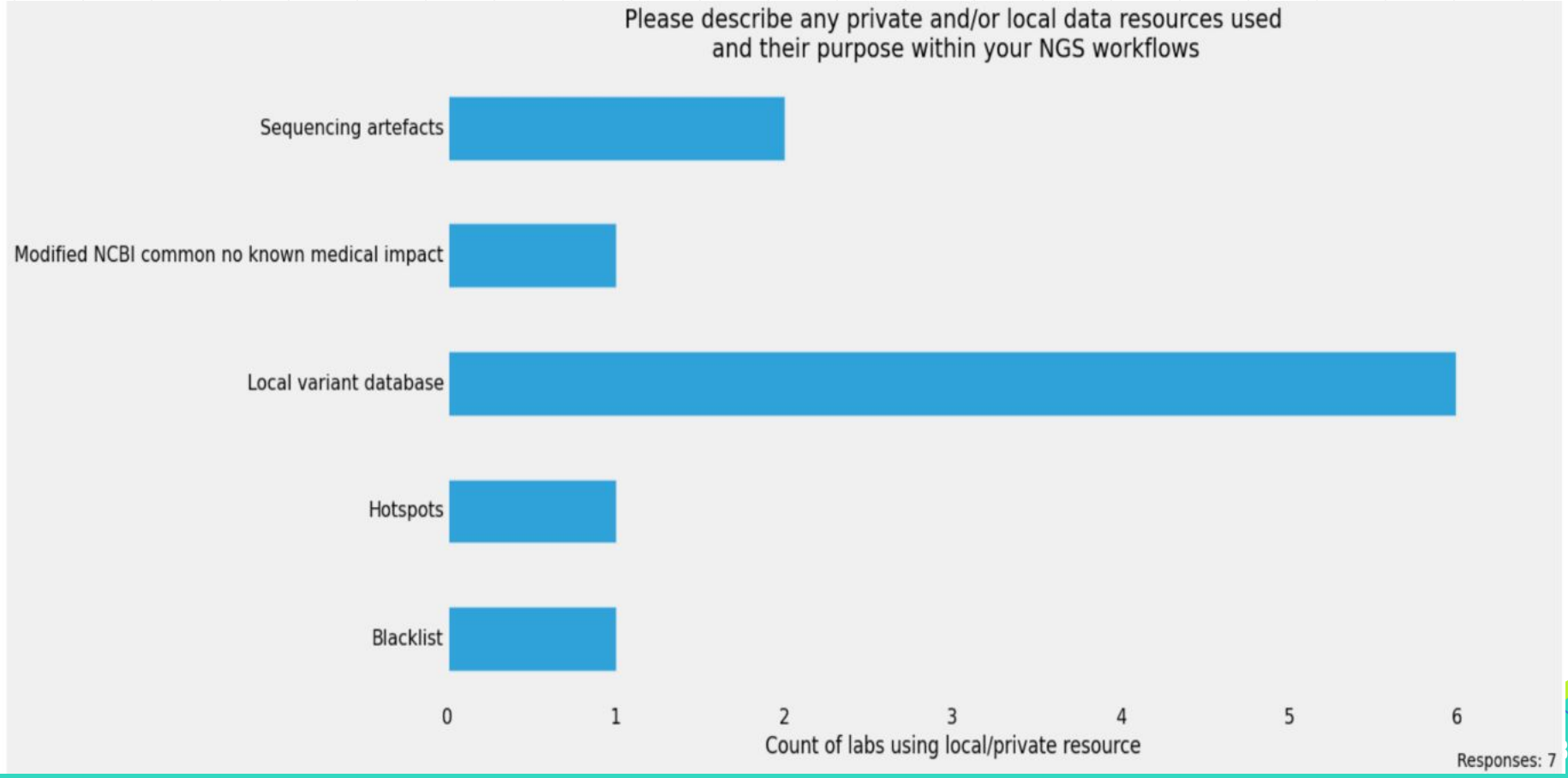
- Most using just the primary sequence
- Limited use of decoy and fix patches
- No use of alt loci



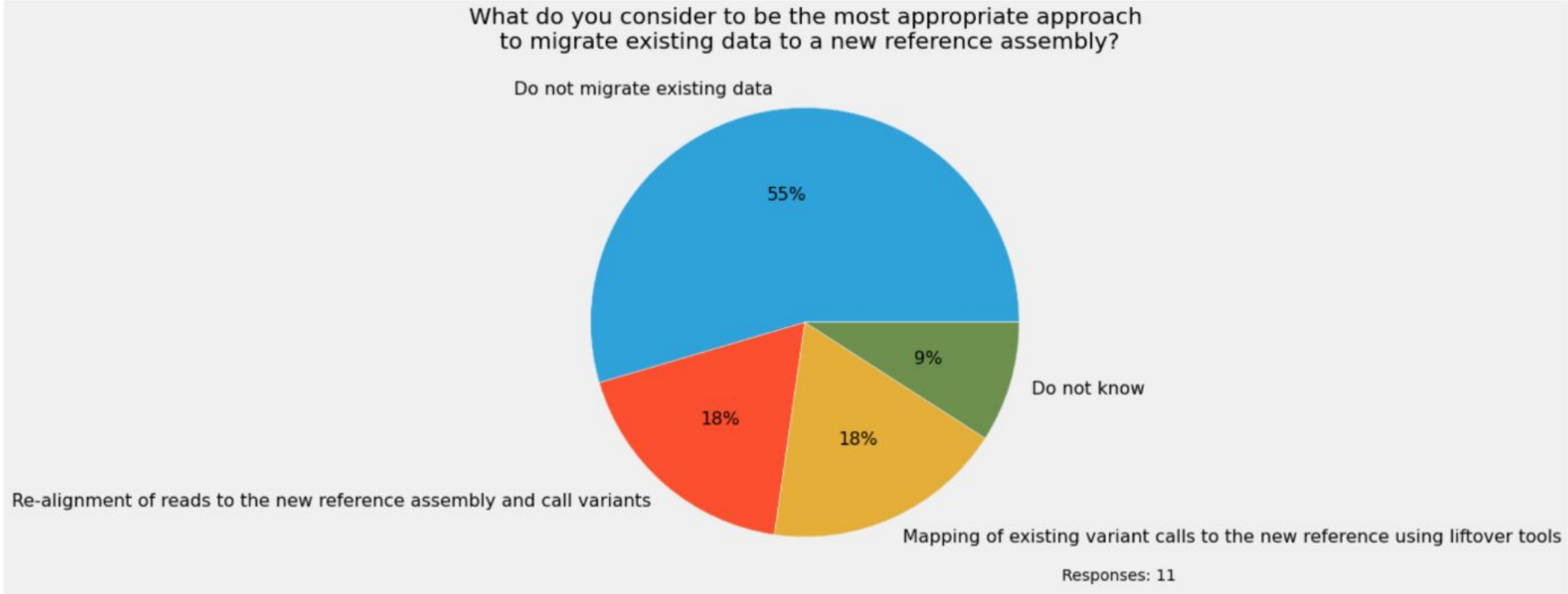
Census survey findings 2



Census survey findings 3

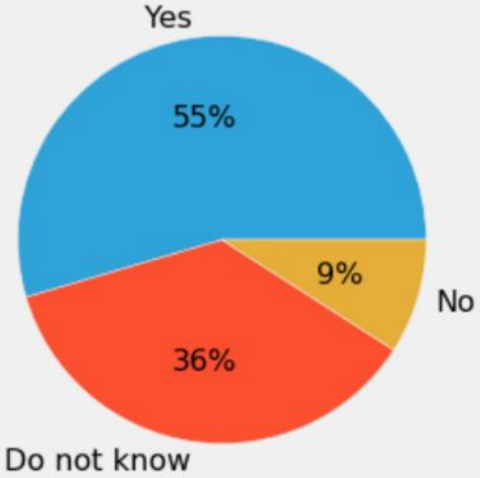


Census survey findings 4

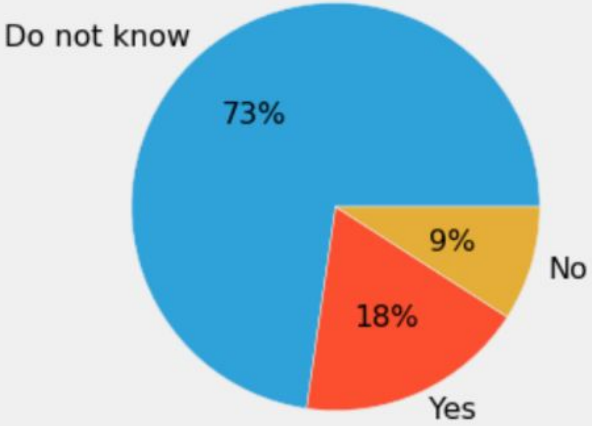


Census survey findings 5

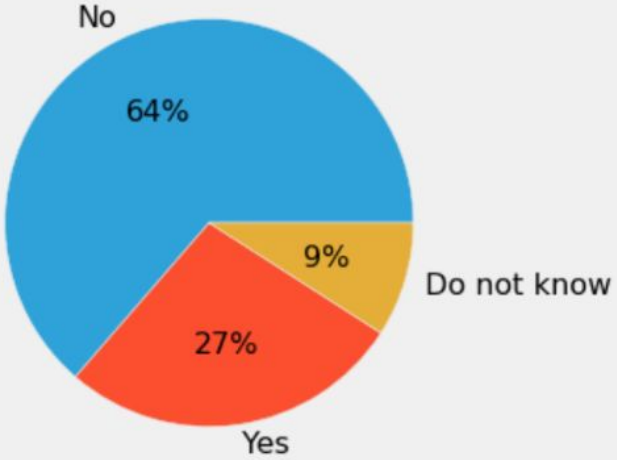
Do you believe migration to GRCh38 could potentially improve the quality of NGS test results generated by the laboratory?



Do you believe migration to GRCh38 could potentially reduce the quality of NGS test results?



Do you consider migration to GRCh38 to be currently feasible for your laboratory?

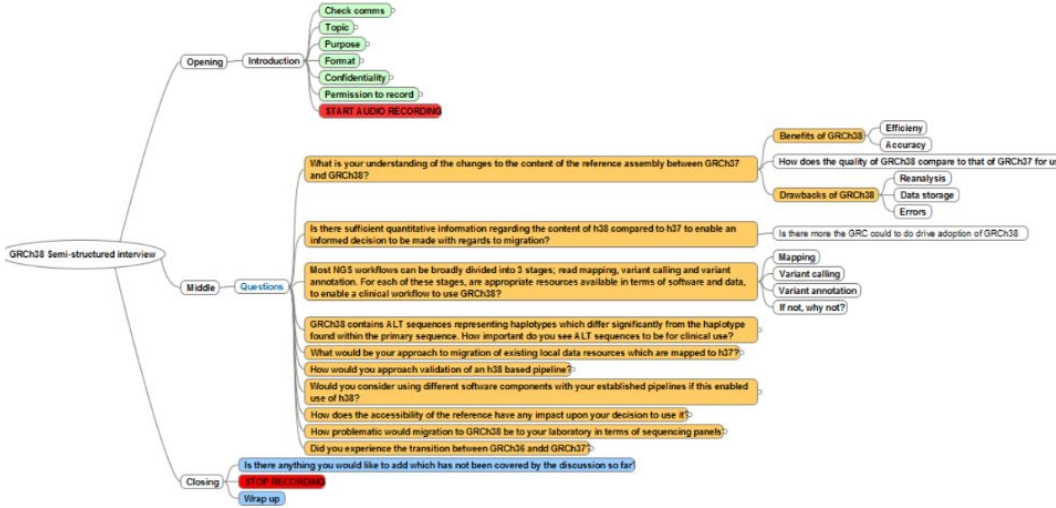


Responses: 11

Semi-structured interviews

16 bioinformaticians invited
6 participated

- Interviews recorded
- Transcribed
- Coded
- Themes identified
- Recoded



Semi-structured interview findings
Motivation and prioritisation

- Implementation of GRCh38 is not a priority
- Migration to the new assembly not investigated in depth

Q: Do you think that there is enough information about the new reference to enable you to make a decision about whether you should migrate?

A: I think there is probably a lot of information out there, I wouldn't say that I actually gone out looking for it, I've just sort of groaned at the prospect, put it on the back burner.

Semi-structured interview findings
Knowledge and awareness

Participants were asked to describe their understanding of the changes to the reference assembly within GRCh38

- None provided a comprehensive description of the changes
- Understood that there are problems with GRCh37 which are addressed in GRCh38
- Detailed understanding of those problems was lacking

Semi-structured interview findings

Expectations

Participants were asked to describe their expectations of the impact of migrating to GRCh38

“Q: What do you see as the benefits of using GRCh38 for clinical testing?

A: I think for most things it's not going to make much difference...for most of the genes on which we're working at the moment which map fine, the transcripts are all fine, I can't see any big difference.”

Semi-structured interview findings
Community

“For me the decision is much more about difficulties relating to confusion between labs...

...the fact is people make the assumption everybody is on 37 at the moment, so you start using 38, there will be mistakes make. There is actually no doubt about that at all.”

(Edited for clarity)

Migration Barriers - technical

- Pipeline components
  - Software
  - Reference files
  - Annotation data
- Pipeline validation resources/datasets
- Migration of existing data
  - Local variant databases
  - Individual reports of variation
  - If/when/how to liftover/remap

Migration Barriers - change implementation

- **Education and Information** – understanding the need to change
  - Understanding of changes between builds is poor
- **Motivation** – wanting to change
  - Limited understanding of benefits of migration
- **Skills** - how to change
  - e.g. liftover between builds
- **Practicalities** – ability to change
  - Availability of resources to needed to implement migration
- **Cooperation** – coordinating change
  - Need to standardisation of approach