Version History

|  |  |  |
| --- | --- | --- |
| Version | Date | Description |
| 0.2 | 2019-03-15 | Version circulated to GLHs for comments |
| 1.0 | 2019-03-29 | Baseline version following GLH comments |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Version 1.0

Contents

[Background 1](#_Toc4756520)

[When is data checked 1](#_Toc4756521)

[Description 2](#_Toc4756522)

[Sequence diagram 1: SNP comparison results in positive match 3](#_Toc4756523)

[Sequence diagram 2: SNP comparison results in negative match 4](#_Toc4756524)

[Sequence diagram 3: SNP comparison finds no GLH SNP 5](#_Toc4756525)

[VCF specification 5](#_Toc4756526)

[Genome version 5](#_Toc4756527)

[The VCF specification 6](#_Toc4756528)

[Minimum requirements 6](#_Toc4756529)

[Dealing with ‘no VCF’ scenario 6](#_Toc4756530)

[Dealing with duplicate entries 7](#_Toc4756531)

[Example message 7](#_Toc4756532)

[Authentication and authorisation 7](#_Toc4756533)

[Exception handling process 7](#_Toc4756534)

# Background

Positive sample identification is required for GMS. This guarantees unequivocally that the genome sequence belongs to a person (whether a diagnosis is found or not) and, for SNVs, obviates the need for orthogonal confirmation of the presence of a variant.

Genomics England, for the 100K project, currently performs what we call genetic vs reported checks. These identify inconsistencies between the clinical data (sex and family relationships) and the sequenced data. These, while helpful, don’t cover all possible swaps (for example same sex parent offspring cases or singletons where the sex matches by chance).

NHS England has requested a mechanism by which the home GLH provides NGIS with a set of SNP genotypes that need to be confirmed against the sequence data. The current proposed solution obtains a copy of the sample prior to sending to the plating lab.

The SNP type submission should eventually be part of messaging between lab and NGIS. Messaging will not be part of MVP

## When is data checked

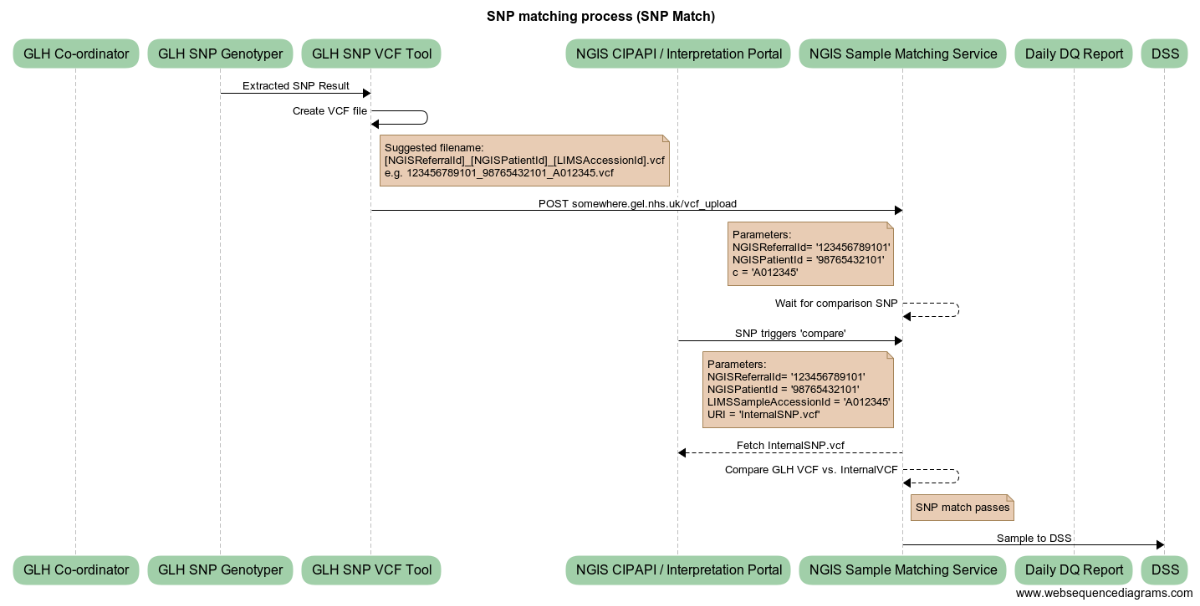
SNP comparison can be triggered by one of two events;

1. The WGS pipeline at GEL creates a SNP profile for each sample; when completed automated interpretation is completed and results loaded into CIPAPI. Upon successful loading into CIPAPI, CIPAPI invokes the comparison.
2. A GLH co-ordinator logs into the interpretation portal and manually invokes a comparison (as part of a resubmission of an erroneous SNP call(s) to correct a failure). The option to programmatically invoke the comparison from the CIPAPI will be available.

# Description

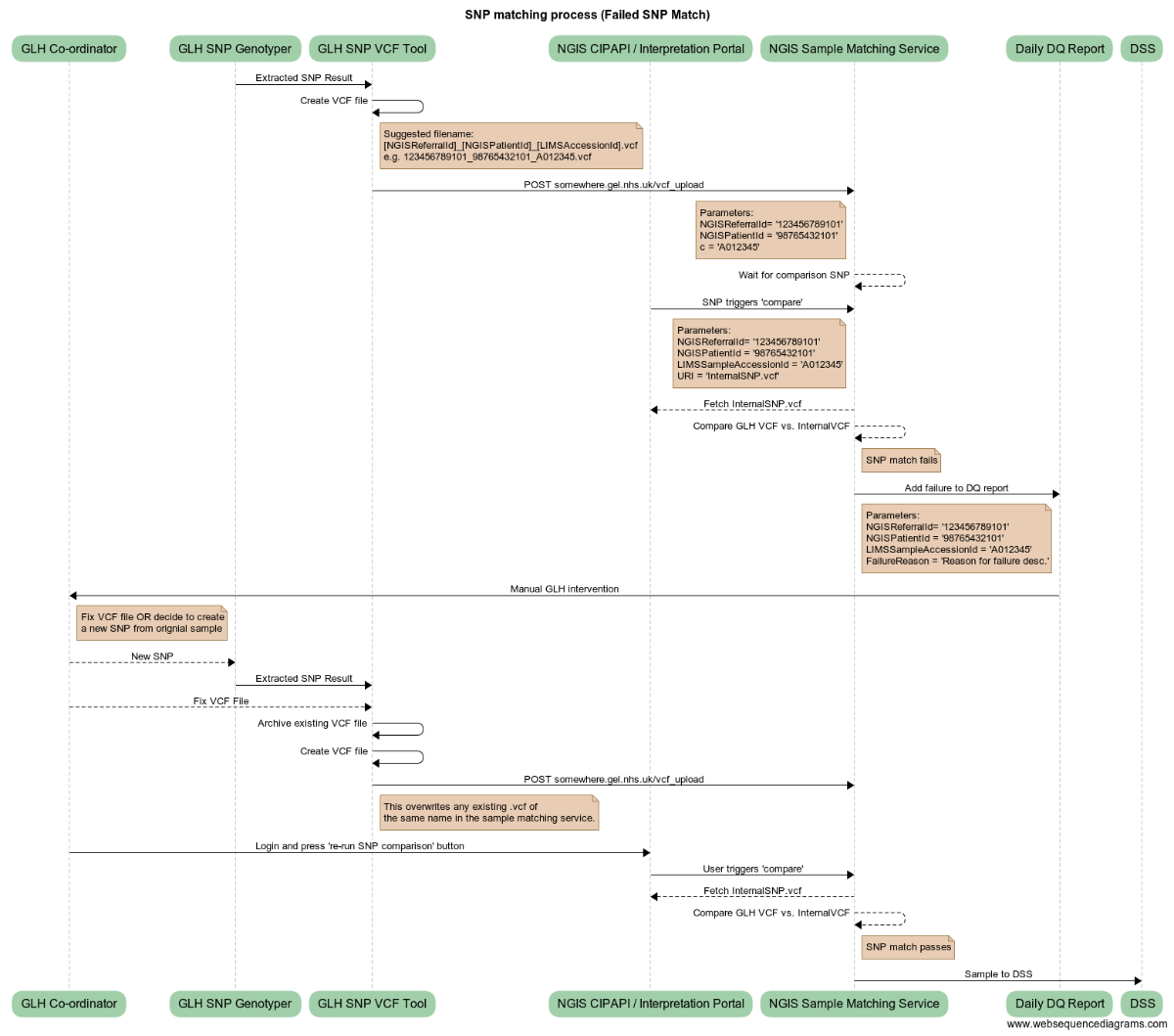
* NGIS will create a service called the ‘NGIS Sample Matching Service’ that is accessible programmatically
* The service allows a GLH to programmatically upload a VCF file with the SNP types.
  + Each VCF represents one sample
  + Each VCF will contain genotype information for ≥24 SNPs, including SNP calls to the reference base.
  + Upon submission of the VCF (via an NGIS secure REST web service) the following parameters will also be included as path parameters:
    - The human readable NGIS Referral ID (12 characters)
    - The human readable NGIS Patient ID (12 characters)
    - The local LIMS Sample Accession ID (string)
* The service supports the following verbs on a VCF:
  + PUT – for adding or replacing a VCF file.
  + GET – for retrieving a previously uploaded VCF file which matches the specified identifiers within the parameters.
  + DELETE – remove a previously uploaded VCF file which matches the specified identifiers within the parameters.
* At present the service will store indefinitely the VCF for a sample. This may be reviewed later.
* The GLH can upload at any time in the process, however if no SNP VCF is, if a GLH SNP VCF is not provided by the time the genome is processed, it may be dispatched to Decision Support System (DSS – currently Congenica) without a check being possible.
* The check will be performed against the version of the genotype file currently stored on the Sample Matching Service at time the comparison was requested. No provision is given for this being updated/deleted immediately afterwards without a corresponding additional comparison request to update reporting.
* GLHs must create a ‘client’ to create and submit VCFs to the API. (Note, to assist GLH’s, GEL can will example code at a later date).
* GLHs may choose to name their local VCF file at their discretion since the service requires the VCF content to be part of the request body. A candidate naming convention is indicated in the sequence diagrams below.
* The Sample Matching Service will confirm that the id used in the VCF’s genotype column ID is consistent with the LIMS Sample Accession ID.
* NGIS will check that the genotypes match between lab and WGS, prior to dispatch to DSS.

## Sequence diagram 1: SNP comparison results in positive match



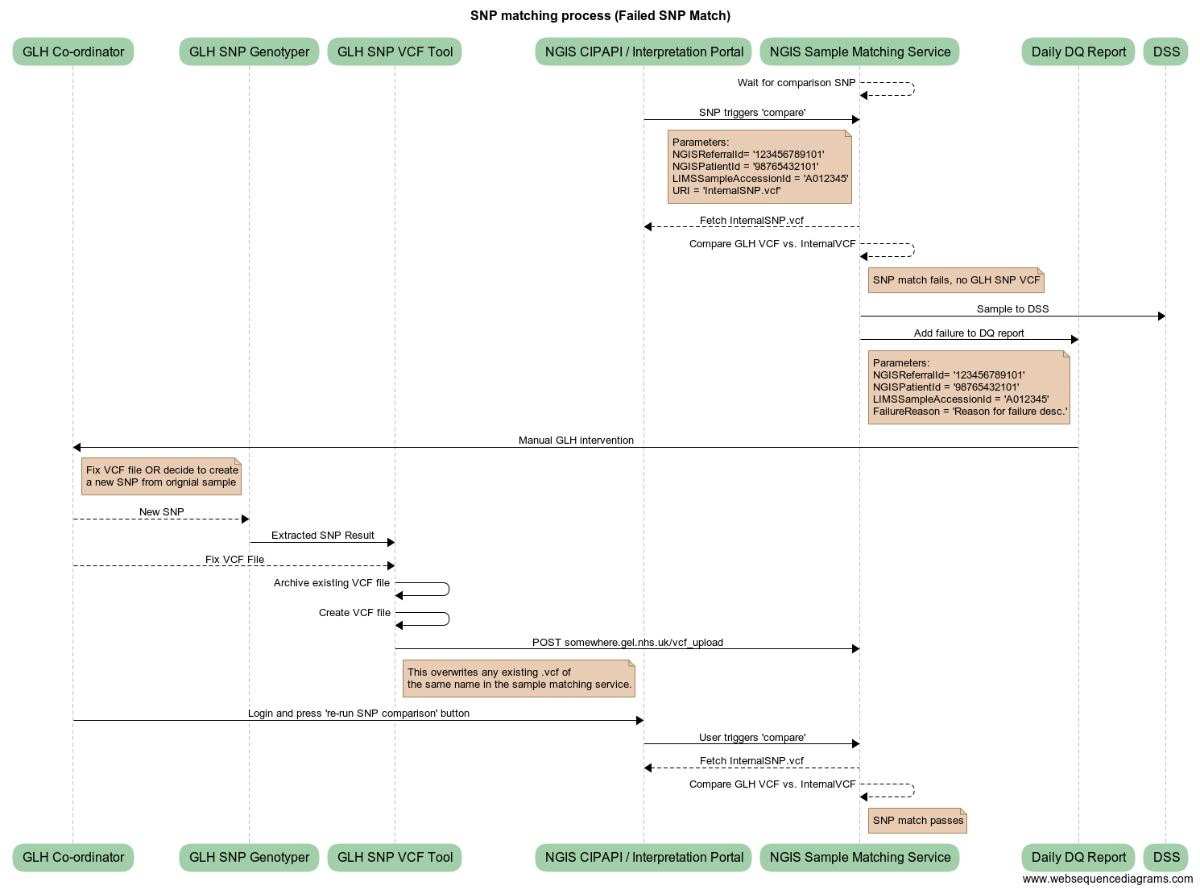
If samples match, then the case is sent to DSS.

## Sequence diagram 2: SNP comparison results in negative match



If samples do not match, then the case is not sent to DSS. Instead, a data quality error is flagged and added to the GLH ‘Daily DQ Report’. Note: Failures will be supplemented with the content from the genetic vs reported checks (sex, family relatedness, Mendelian inconsistencies)

## Sequence diagram 3: SNP comparison finds no GLH SNP



If SNP types are not available the sample is “pending” and proceeds to DSS.

**In case that genotypes are submitted after the sample has been dispatched to the DSS and the check fails, it remains the responsibility of the plating GLH to alert the lab performing the interpretation.**

# VCF specification

## Genome version

Variants must be referenced to the GRCh38 genome version.

It is expected that the GLH represents their alleles and positions using the same reference file that Genomics England uses: <http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/README.20150309.GRCh38_full_analysis_set_plus_decoy_hla>

Chromosome names should be prefixed with "chr". The mitochondrion is called "chrM". There following values are permitted: {chr1, chr2, chr3, chr4, chr5, ..., chr21, chr22, chrX, chrY, chrM}.

## The VCF specification

Files will be submitted using the VCF 4.2 specification: <https://samtools.github.io/hts-specs/VCFv4.2.pdf>

VCF files must only contain one sample.

## Minimum requirements

The header shall include the following tags:

##fileformat=VCFv4.2

##fileDate=**<date the file was produced: eg. 20090805>**

##source=**<data source: e.g. mygenotypingplatformV3.4>**

##reference=**<name of the fasta file used>**

##contig=**<as per the specification>**

* The *source* field must be from an enumeration reflecting the standardised names of the assays used for genotyping.
* The genotype field (column) must be named by the LIMS Sample ID (*primary\_sample\_id\_in\_glh\_lims* in GEL1001/TOMS).
* Genotypes must contain the *GT* field.
* It is highly desirable, **but not required,** to include the Genotype Likelihood *PL* field where possible. This allows a more sophisticated concordance check.
* It is desirable, **but not required,** to add a genotype quality *GQ*
* Only biallelic SNVs are permitted. The *ALT* field must not contain more than one value
* Only *PASS* variants will be used
* Variants should be normalised, i.e. parsimonious and left-aligned
* Variants should be sorted by reference contig name and position.

## Dealing with ‘no VCF’ scenario

* TBC: potentially a new endpoint will be written to support this.

## Dealing with duplicate entries

* If genotypes for the same sample are re-submitted either on purpose or by mistake, the latest genotype according will overwrite the previously stored variant calls.

## Example message

|  |
| --- |
| ##fileformat=VCFv4.2  ##fileDate=[File\_Date]  ##source=my\_assay  ##reference=http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38\_reference\_genome/GRCh38\_full\_analysis\_set\_plus\_decoy\_hla.fa  ##contig=<ID=chr20,length=62435964,assembly=GRCH38,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>  ##FILTER=<ID=PASS,Description="All filters passed">  ##FILTER=<ID=NOCALL,Description="Genotype not called on array.">  ##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">  ##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">  #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT **NA00001**  chr20 14370 rs6054257 G A 29 PASS . GT:GQ 0|0:48 1|0:48  chr20 17330 . T A 3 PASS . GT:GQ 0|0:49 |

**Note that for MVP only single sample VCFs will be accepted.**

# Authentication and authorisation

The service will sit behind the KONG API layer and is therefore proposed to be authenticated using the same mechanism as existing CIPAPI’s.  
  
Note: This approach is being confirmed but should not materially change the GLH’s from building against an unauthenticated test end point initially.

# Exception handling process

Any validation errors, processing failures, or SNP comparison failures (including no GLH SNP) are added to the daily DQ report and are not removed until resolved.

A GLH Co-ordinator must work to resolve any failures which may include creating a JIRA ticket with the GEL service desk if the issue is found to be outside of the SNP generation and VCF uploading process (for example, if the genome sequence needs manual intervention at the DSS stage).

An NHS England SOP/policy will include details of what action to take in various failure scenarios beyond those remediable by correcting the SNP VCF sample.