Small Variant PON Improvements

Background

PON filtering is performed to filter out both common artefacts and likely germline variants. Although genuine somatic variants may occur at sites which fall into one of these categories, we have tolerated filtering these as variants which are common in the germline are hence unlikely to be pathogenic. Furthermore, there are relatively few artefact sites in our panel regions, and hence calls at artefact locations tend to be overwhelmingly artefacts.

We filter based on the following 'PONs':

- GNOMAD (AF>0.00015)
- WGS PON (based on 1000 samples in 37, filter thresholds are tier dependent)
- PON_ARTEFACT (a panel specific set of artefacts. Always filter if overlaps)

We see 2 major issues with current PON functionality

1. Our various PONs + GNOMAD may overlap sites which are annotated as known or likely pathogenic sites in clinical annotation databases, particularly clinvar and our germline/somatic HOTSPOT lists. These currently get PON filtered (and hard filtered in tumor only) despite their pathogenic status)
2. More generally we sometimes see candidate variants at common artefactual sites which appear clearly genuine due to the AD/AF being much higher than the artefact observed distribution.

Changes in WiGiTs v2.0

Remove Pathogenic/Likely Pathogenic variants from 37 & 38 PONs

Remove variants from existing WGS PON and artefact PONs if they meet the following criteria:

- PON_COUNT < 10
- Germline/somatic HOTSPOT or CLINVAR pathogenic
- RepeatCount < 4 (this is only checked for indels)

These rules will also be applied in the future whenever new WGS or Artefact PONs are created. The specific variants impacted can be found here.

CLINVAR pathogenic variants are limited to within exons or splice regions for canonical transcripts of genes within driver panel.

We also follow the existing hmftools convention for determining pathogenecity from CLINVAR annotations – namely CLNSIG contains Pathogenic/Likely_pathogenic, or if conflicting, the CLNSIGCONF contains Pathogenic/Likely_pathogenic without containing Benign/Likely_benign.

Do not PON_GNOMAD filter pathogenic variants if GNOMAD_FREQ<1%

Germline/somatic HOTSPOT or CLINVAR pathogenic will no longer be filtered by GNOMAD if population frequency is relatively low (ie GNOMAD_FREQ < 1%)

Conditionally recover artefact-prone pathogenic variants with high VAF

Germline/somatic HOTSPOT or CLINVAR pathogenic variants will no longer be filtered by WGS PON if mean PON reads < 6 and VAF > 8%. If the variant is an indel in a microsatellite context spanning N ref bases, we also require VAF > N%.

Summary of new behaviour

WGS PON

- If germline/somatic HOTSPOT/CLINVAR pathogenic with PON_COUNT < 10 and RepeatCount < 4, pass
- Otherwise, if germline/somatic HOTSPOT/CLINVAR pathogenic with mean PON reads < 6 and VAF > max(0.08, 0.01*numRepeatBases), pass
- Otherwise filter if the following is satisfied based on variant TIER:

TIER	Filter if...
HOTSPOT	PON_MAX _READS >= 5 & PON_COUNT >= 10
PANEL	PON_MAX_READS >=5 & PON_COUNT >= 6
OTHER	PON_COUNT >= 6

(note: this assumes the hg38 WGS PON is regenerated on 1000 samples)

ARTEFACT PON (NOT APPLIED IN WGS)

- If germline/somatic HOTSPOT/CLINVAR pathogenic with PON_COUNT < 10 and RepeatCount < 4, pass
- Otherwise, filter

GNOMAD

- If germline/somatic HOTSPOT/CLINVAR pathogenic, filter if frequency > 1%
- Otherwise, filter if frequency > 0.015%

Future Improvements

- Move from AD to VAF base counting in PON annotations

- With improved annotations, make a classifier distinguishing ‘likely artefactual’ PON variants from ‘likely germline’ ones
- Extend conditional variant recovery based on sample-specific VAF to variants that are not germline/somatic HOTSPOT/CLINVAR pathogenic
    - This will probably be a function of typical PON VAF, repeat count (for an indel), and also not apply to variants classified as ‘likely germline’
    - Analysis has been done on this, but it remains out of scope for v6 due to the limitations of current PON annotations, and given we expect big changes to Sage variant calling + PON counts in the next release
    - Meanwhile, impact from germline/somatic HOTSPOT/CLINVAR pathogenic sites is much more limited, and each variant recovered has been individually assessed for ~500 WGS samples
- Ideally generate a PON with VAFs from Peter Mac Panel data