Mutation short label annotation rules adaptation to HGVS rules

Rules used for short label notation in mutation features adapted to HGSV rules, which can be found here: http://varnomen.hgvs.org/recommendations/protein/**. We always add the UniProt accession of the affected protein as a prefix (e.g. P12345:p.Ala34Cys).**

- Single amino acid change:
    - Old rules: Three-letter code of original aa is used, then sequence position, then three-letter code of replacement (no capitals).
        - Examples:
            - ile345thr
    - HGVS: Is considered a ‘substitution’, defined as “*a sequence change where, compared to a reference sequence, one nucleotide is replaced by one other nucleotide*”. Amino acids need to be depicted using the three-letter code, using capitals.
        - Format: *“prefix”“amino_acid”“position”“new_amino_acid”*
        - Examples:
            - P12345:p.Ile345Thr
        - For test: EBI-9095885, EBI-8524086, EBI-9825301

- Multiple amino acid change, non-sequential positions:
    - Old rules: Same rule as above, comma-separated.
        - Examples:
            - ile234ala,thr345ala,pro456ala
            - ile345ala,ile678ala,ile897ala
    - HGVS: Is considered as multiple substitutions, which can be grouped using brackets and separated using semicolon.
        - Format: *“prefix”[“amino_acid$_1$”“position$_1$”“new_amino_acid$_1$”, “amino_acid$_2$”“position$_2$”“new_amino_acid$_2$”,…]*
        - Examples:
            - P12345:p.[Ile234Ala;Thr345Ala;Pro456Ala]
            - P12345:p.[Ile345Ala;Ile678Ala;Ile897Ala]
        - For test: EBI-9693147, EBI-10889784, EBI-15731927

- Multiple amino acid change, sequential positions:
    - Old rules: Original aas separated by underscore, then sequence range separated by hyphen, then substitution amino acids separated by underscore. Replacement residues are not collapsed even if they correspond to the same amino acid (all changed to alanine, for example).
        - Examples:
            - pro_pro_pro_pro123-126ala_ala_ala_ala
            - pro_thr_leu12-14ala_ala_pro
            - thr_ile_cys_tyr30-33ala_ala_ala_ala
    - HGVS: It is considered a deletion-insertion case, defined as “*a sequence change where, compared to a reference sequence, one or more amino acids are replaced with one or more other amino acids and which is not a substitution or conversion*”.
        - Format: *“prefix”“amino_acid(s)+position(s)_deleted”“delins”“inserted_sequence”*
        - Examples:

            - P12345:p.Pro123_Pro126delinsAlaAlaAlaAla
            - P12345:p.Pro12_Leu14delinsAlaAlaPro
            - P12345:p.Thr30_Tyr33delinsAlaAlaAlaAla
            - P12345:p.Cys28_Lys29delinsTrp
            - P12345:p.[Pro578_Lys579delinsLeuTer]

        - For test: EBI-11178974, EBI-11314033, EBI-12590047, EBI-8839684, EBI-9846491, EBI-2891626, EBI-15582875

- Deletions:
    - Old rules: Deletions are represented with the three-letter code **"del"**. For representation of deletions in the *"resulting sequence"* field dots are used: **"."**. NOTE: It is VERY IMPORTANT to add the dots in the *"resulting sequence"* field, so we can be sure there is no missing information in the field and a deletion is intended to be represented. When the deleted region spans over three residues in length, it is not a mutation anymore, but a region affecting binding, and should be represented with another feature type: *"required to bind region"*.
        - Examples:
            - pro123del
            - pro345del,ile347del
            - leu_leu_leu356-358del_del_del
            - arg_pro_arg_lys_arg123-127del_pro_del_lys_del
    - HGVS: A ‘deletion’ is defined as “*a sequence change where, compared to a reference sequence, one or more amino acids are not present (deleted)*”.
        - Format: *“prefix”“amino_acid(s)+position(s)_deleted”“del”*
        - Examples:
            - P12345:p.Pro123del
            - P12345:p.[Pro345del;Ile347del]
            - P12345:p.Leu356_Leu358del
            - P12345:p.[Arg123del;Arg125del;Arg127del]
        - For test: EBI-6898602, EBI-16008622, EBI-9085688, EBI-1641252
    - Note: ‘Required to bind’ annotations could also be re-labelled using this rule, unless a stop codon is introduced at the end of the sequence. Example: p.Trp126Ter or p.Trp126*.

- Insertions:
    - Old rules: Same rules applied for multiple amino acid changes in sequential positions are used, adding replacement amino acids in the second part of the expression. If there is no substitution of original residues, they are included in the replacement part of the statement.
        - Examples:
            - cys_thr123-124cys_pro_ala_thr - *Cys and thr do not get replaced, there is only an insertion of pro and ala between them.*
            - thr_lys12-13thr_pro_pro_lys - A *couple of prolines get inserted between thr and lys.*
            - HGVS: An ‘insertion’ is defined as “*a sequence change where, compared to the reference sequence, one or more amino acids are inserted and where the insertion is not a copy of a sequence immediately N-terminal (5')*”. Both flanking amino acids are required to generate the annotation.
        - Format: *“prefix”“amino_acids+positions_flanking”“ins”“inserted_sequence”*
        - Examples:
            - P12345:p.Cys123_Thr124insProAla

            - P12345:p.Thr12_Lys13insProPro

        - For test: EBI-2891626, EBI-11475055

        - SPECIAL CASE: N/C-terminal insertion (examples: EBI-16879916, EBI-5260369).

            - These are extensions in HGVS. They have very specific cases covered:
                - p.Met1ext-5: a variant in the 5' UTR activates a new upstream translation initiation site starting with amino acid Met-5. Resulting sequence can be reported as such: MXXXM.
                - p.Met1_Leu2insArgSerThrVal: amino acid Met1 is changed to Val activating an upstream translation initiation site at position -4 (Met-4), insertion amino acids ArgSerThrVal between Mat1 and Leu2. NOTE: this variant is not described as an extension (p.Met1Valext-4) since Met1, part of the normal amino acid sequence, is changed. Resulting sequence would be MRSTVL.

            - C-terminal cases would need to be added with range "c-c". Only additional amino acids would need to be stated in the resulting sequence field.
                - P12345:p.Ter110Glnext*17: a variant in the stop codon (Ter/*) at position 110, changing it to a Gln-codon (a no-stop variant) and adding a tail of new amino acids to the protein's C-terminus, ending at a new stop codon 17 residues down the old one.
                - P12345:p.Ter315TyrextAsnLysGlyThrTer: a variant in the stop codon (Ter/*) at position 315, changing it to a Tyr-codon (a no-stop variant) and adding a tail of new amino acids to the protein's C-terminus, ending at a new stop codon.
                - P12345:p.Ter327Argext*?: a variant in the stop codon (Ter/*) at position 327, changing it to an Arg-codon and adding a tail of new amino acids of unknown length (position *?) since the shifted frame does not contain a new stop codon.
            - At the moment, previous cases are handled as deletion-insertions:
                - P12345:p.Arg123delinsArgThrPro: TP are added after R, which is the last amino acid in the original sequence.
                - Alternative: P12345:p.Ter124ThrextProTer. Range would be c-c and resulting sequence TP.

    - Handling insertions in the mutations update:

        - Both flanking positions need to be given, so an insertion will always need to have a range with consecutive start and end positions.

        - If an insertion is identified but start and end position are the same, the mutation update will throw a message and identify these cases in a special section, so they are checked and corrected manually.

- Unconventional / unknown amino acid substitutions:

  - Old rules: Instances where the resulting amino acid is unknown or an unconventional amino acid is used (e.g. pBpa, crosslinkable amino acid p-benzoyl-L-phenylalanine) use the three-letter code **"xaa"**, represented with the one letter code **"X"** in the *"resulting sequence"* field.
  - HGSV: Just use the IUPAC convention and capitalize the first letter of the three-letter code: “**Xaa**”.
      - P12345:p.Leu286Xaa

- Poly-Qs and repetitions:
  - Our rules for polyQs: We use the notation ‘*q**N** poly-glutamine extension*’, where ‘***N***’ is the number of glutamines.
    - Example: q97 poly-glutamine extension
  - HGVS: It is considered a ‘repeated sequence’, defined as ‘*a sequence where, compared to a reference sequence, a segment of one or more amino acids (the repeat unit) is present several times, one after the other*’.
    - Format: *“prefix””amino_acid(s)+position_repeat_unit””[””copy_number””]”*
    - Examples:
      - P12345:p.Gln18[97]
      - P12345:p.Gln24[35_40] *(used if the amount of Qs is not fully known and ranges between 35 and 40)*.

      - P12345:p.Ala23[4]: original seq A23 -> resulting seq AAAA23

      - P12345:p.Ala23_Pro25[3]: original seq AHP(23-25) -> resulting seq AHPAHPAHP23

- Frame shifts:

  - Until now, we could not report frameshifts. We used required to bind features in the cases where introducing a premature stop codon in one of the participants would prevent an interaction. With HGVS rules, we can potentially annotate these cases as mutations. We have curated one case: check EBI-20559947 in the editor.

    - Format: *“prefix””amino_acid”position”new_amino_acid””fs””Ter””position_termination_site”*
    - Examples:
      - P12345:p.Arg123LysfsTer34. Resulting sequence would be LXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX*