Beyond the Exome - P5
Variant information in non-coding regions
Annotation Guidelines (v1.0)

Samuele Garda
gardasam@informatik.hu-berlin.de

1 Introduction

This document describes the guidelines used for the construction of the annotations of several biological elements in PubMed abstracts. The aim is to create a corpus with human-curated annotations specifically developed to investigate variants in non-coding regions of the genome and their effects.

The annotation process will focus on *things*, such as genes, diseases and genomic positions. These are referred to as *entities*. The annotation process itself shall rely on (1) common sense, (2) domain background knowledge and (3) consultation of external resources.

Whenever possible/relevant an annotation example is provided. In these examples a span of text with a green background means the the annotation is correct, whereas a red background means the annotation is incorrect. Finally, the background yellow means that a span of text is missing from the annotation.

The document is structured as follows:

- in Section 2: general guidelines to be followed throughout the annotation process;
- in Section 3: guidelines specific to each entity:
  - GENE (Section 3.1)
  - DISEASE (Section 3.2)
  - TISSUE (Section 3.3)
  - SPECIES (Section 3.4)
  - VARIANT (Section 3.5)
  - REGION:
    - ENHANCER
    - PROMOTER
    - TFBS (Section 3.6.1)
- in Section 4: how each entity should be normalized;

2 General

2.1 Annotation process

This section contains general advice for the annotators when performing the annotation task.

1. CONSISTENCY: try to come up with a coherent process on how to determine entity boundaries and normalization;
2. NORMALIZATION: Please try to normalize the entity, e.g. gene "epsilon globin" should be mapped to the NCBI Entrez Gene ID 3046;
3. EXTERNAL RESOURCES: you can (and should) consult knowledge resources (e.g. Wikipedia, NCBI, OMIM,...);
4. COMMENTS: whenever an entity is ambiguous, **PLEASE ADD A COMMENT** in the annotation. This helps the discussion;

2.2 All entities

This section contains annotation guidelines that holds for any type of entity.

1. EXPLICIT MENTION ONLY: only entities that have an explicit mention in the text should be annotated. General terms must be avoided, e.g.
   - The binding of some transcription factors (TFs) is diminished by methylation of their target sequences.
   - FOXP2 in 7q31 cause a rare disorder involving
   - Adiponectin (AdipoQ) expression inversely correlates
2. COMPLETENESS: annotations should respect semantic constraints, e.g.
   - in their phenotypic expression of IgG receptors
   - in their phenotypic expression of IgG receptors
3. EXHAUSTIVENESS: all the entities in the document must be annotated, regardless if they are mentioned once or multiple times;
4. SPECIAL CASES:
   - ABBREVIATIONS/ACRONYMS should be marked consistently and include the entire corresponding text, e.g.
     - levels from lymphoblastoid cell lines (LCLs) identified
     - ERG in a human erythroleukemia cell line , HEL
   - MULTI-WORD expressions must be annotated **entirely**. If present, the abbreviations must be annotated with all the relevant entities, e.g.
     - a dual promoter-enhancer (DPE) controls (# both marked both as ENHANCER and PROMOTER)
     - a dual promoter-enhancer (DPE) controls

3 Entities

This section reports annotation guidelines and examples specific for each entity of interest.

3.1 GENE

MUST be annotated as GENE:

- Transcription Factor, mRNA, protein, or enzymes;
- two-word entities where one of the words is an abbreviation and the other word is a number or single letter (e.g. "Cox 2", "H Ras", "miR 145");
- continuous string of characters (e.g. "BRCA 1 or 2"). The continuous string shall be marked and normalized **separately**,e.g.
  - Rab1B, -5, -7, -8, or -11A
- identifiers and/or accession numbers from e.g. UniProt, RefSeq, Ensembl;
- outdated and non-standard symbols, e.g.
  - its expression significantly decreases GnRH gene expression

MUST NOT be considered GENE mentions:

- group of genes or complexes, e.g.:
  - overproduction of type I collagen and development
  - Wnt proteins are secreted cytokines and several
  - but defects in SWI/SNF targeting and control of enhancer activity
- elements too broad
  - in part through inhibiting p38 MAPK, which resulted in a decrease

3.1.1 Notes:

General genes/proteins concepts in form of isolated term ("gene", "proteins", "peptide",...) shall be excluded from the annotation, unless they are strictly necessary to the identification of the entity, e.g.

- We sought to exploit these two diagnostic SNPs located 168 bp apart on gene phoE
- We sought to exploit these two diagnostic SNPs located 168 bp apart on gene phoE
- the G gamma gene in embryonic cells and the beta gene in fetal and adult erythroid cells

3.2 DISEASE

MUST be considered DISEASE mentions:

- medical conditions in their complete form (the entire span of text necessary to identify the condition), e.g.
  - child with severe speech and language disorder separates FOXP2
  - child with severe speech and language disorder separates FOXP2
  - child with severe speech and language disorder separates FOXP2
- non-specific medical conditions, e.g. mention of disease categories:
  - band potentially other cancer types.
- general mentions of "tumor" (it does not necessarily imply disease), e.g.
  - multiple common tumor types, including a melanoma
  - many types of solid tumours including the lethal castration-resistant prostate cancer (CRPC)

MUST NOT be considered DISEASE mentions:

- modifiers
  - Histone Deacetylase Inhibitors Repress Tumoral Expression of the Proinvasive Factor RUNX2
  - targeting strategy to counteract the oncongenic function of RUNX2 in cancer cells

3.3 TISSUE

MUST be considered TISSUE mentions:

- mentions of general tissues, e.g.
  - through the activation of liver X receptor and retinoid X receptor
  - many aspects of neural development
- mentions of specific cell lines, e.g.
  - Wnt3a secreted from human embryonic kidney (HEK)-293T cell
  - The target genes identified in Ramos and BEAS2B cells in this study
  - expression in both prostate cancer cell lines and cell lines of liver origin
  - in the normal lung epithelial cell line BEAS2B
  - was analyzed in beta cell and non-beta cell lines
- even if name contains a DISEASE or a GENE, e.g.
  - such as systemic sclerosis (SSc) [· · ·] In SSc dermal fibroblasts , increased mTOR/Akt signaling
  - in the Burkitt's lymphoma cell line
  - This served to isolate MesP1-positive cells

MUST NOT be considered TISSUE mentions:

- text referring to cancer (as it is labeled as DISEASE), e.g.
  - in patient-derived HER2+ breast carcinomas compared

3.3.1 Notes:

The word "tissue" should not be included in the entity, e.g.

- in response to a decrease in oxygen availability in kidney tissue .

3.4 SPECIES

MUST be considered SPECIES mentions:

- laymen terms for species, e.g. "human", "rat"
- even if they are used as modifiers, e.g.
  - overexpressed at both mRNA and protein levels in human prostate cancer
- scientific terms for speceis, e.g. "Mus musculus"
- any taxonomic rank, e.g.
  - which are well conserved across many mammalian species

MUST NOT be considered SPECIES mentions:

- ambiguous or only implicit reference to a species, e.g. in the example below "child" does not necessarily refer to a human:
  - A recent report described a child with speech and language deficits

3.4.1 Notes:

The annotation should include only the species word(s) without additional terms, e.g.

- in the transgenic mice, WT1 protein expression

3.5 VARIANT

MUST be annotated as VARIANT:

- Human Genome Variation Society (HGVS)-like notation e.g. 111:987654G>A
- dbSNP ID (e.g. rs0000000)
- Genbank accession number, e.g. "NM_006516"
- variations described as distance from the transcription start site (TSS), e.g. "80 bp upstream of the TSS"
- variations described as distance from the start ATG (first coding base), e.g. "c.-390delA" (390 bp upstream of the start ATG)

3.5.1 Notes:

In the case of dbSNP IDs, exclusively the ID must be annotated, e.g.

- assess the association of the rs33912345:c.421A>C and rs10483727 variants
- assess the association of the rs33912345 :c.421A>C and rs10483727 variants

In all other cases the annotation should include all the information provided, e.g. including "g" signaling genome vs transcribed

- DNA sequence variants (DSVs), g.17483576C>G and g.17483564C>T, were identified
- heterozygous DSV, g.17483557Ins was found

3.6 REGION

This macro-category encompasses DNA sequences which are of special interest for the project. These are typically non-coding regions, i.e. different from genes. The regions of interest are *cis-regulatory elements* and are further classified into the more fine-grained entities: TFBS, PROMOTER, and ENHANCER.

For these entities the EXPLICIT MENTION ONLY principle must be followed strictly, i.e. these entities must be used **only for mentions of regions to which - in principle - it could be possible to assign genomic coordinates**.

WARNING : If the regions are described in terms of their proximity to a gene, the REGION entity shall prevail over the GENE one if it is evident that the gene name is used only for identifying the region. In such cases the entire text span referring to the region **must** be annotated, e.g.

- the mammalian GnRH promoter has been extensive studied
- the mouse Pdx-1 gene promoter (2.7 kb) was analyzed
- genomic deletions that included a distant FOXF transcriptional enhancer mapping

In the case other type of entities (e.g. TISSUE, SPECIES, ...) are used to define the REGION mention, the REGION entity prevails if and only if it is used consistently throughout the document, e.g.

- PMID-25486239
  - hypothalamic arcuate nucleus-specific enhancer region of Kiss1 gene in mice [· · ·]
  - study aimed to determine ARC-specific enhancer(s) of the Kiss1 gene by [· · ·]
  - unctions as an enhancer for ARC Kiss1 gene expression in mice

WARNING : genomic coordinates referring to a REGION mention must be annotated as well, e.g.

- the previously described PDX-1-binding site (-2721/-2646 bp)
- E-box (-104/-99 bp) in the proximal region
- uncover an enhancer binding site located within the -150 to -130 bp region
- highly conserved SIX6 enhancer (Ch14:60974427-60974430) were sequenced
- a site located between nucleotides -2,113 and -1,778 . This fragment features enhancer activity
- cardiovascular stem cell using the 3.4-kb MesP1 promoter fragment

3.6.1 TFBS

MUST be annotated as TFBS:

- explicit mentions, e.g.
  - Using ChIP Seq, we identified 556 and 467 putative STAT6 target site .
  - our studies suggest that the putative -45 to -39 MAZ-binding site regulates the constitutive activity
  - the octamer sequence 5'-ATGCAAAT-3' at -10.2 kb
- in the context of genes/genome, typically expressions containing the words:
  - "motif", e.g.
    - Targeted genomic deletion (¡7 bp) of the MITF motif within

    * a far-upstream functional Oct-1 enhancer motif
- "response element", e.g.
    * the core enhancer element derived from the androgen response element III (ARE III)
    * the Il12b enhancer contains functional ISREs for recognition
- "consensus sequence", e.g.
    * transactivation at the sense promoter by binding to a PAX5 consensus sequence
- multiple TFs and TFs families
    * Our data provide significant insight into the functional characteristics of YY1 binding sites

4 Normalization

This section in Table 1 presents the reference ontologies used to normalize the annotations.

ENTITY	ONTOLOGY/FORMAT	EXAMPLE
GENE	NCBI Entrez Gene	"HBE1": 3046
REGION: • ENHANCER • PROMOTER • TFBS	BED (CHR,START,END)	1,1000,1500 (Chr 1 1000-1500 bp)
VARIANT	dbSNP or VCF (CHR,POS,REF,ALT)	• rs823294 • 7,4545,A,C (for SNVs) • 17,3232,AC,A (for InDels)
DISEASE	-	-
SPECIES	NCBI Taxonomy	"human": 9606
TISSUE	Brenda Tissue Ontology (BTO)	"HEK-293A cell": BTO_0002524

Table 1: Entities and their normalization nomenclatures.

4.1 GENE

WARNING: If the GENE symbol is misspelled, i.e. it is written in a format the is not correct according to the context, e.g. a mouse gene written in the human format (1) or the rat form instead of the mouse one (2), e.g.

1. ... livers from Prep1 (i/i) mice feature increased sensitivity to insulin, ... Prep1 overexpression in HepG2 liver cells upregulated SYP and SHP1
2. ... the mouse Pdx-1 gene promoter (2.7 kb) was analyzed ... activation was a prerequisite for SREBP-1c suppression

The **normalization should always follow the context**, e.g. if the sentence is about mouse genes all mentions must be normalized to the mouse IDs, even if for instance are written in the human format.

If it is not possible to infer from the context the species, the gene normalization should be done according to its symbol.

4.2 REGION

WARNING: REGION entities (ENHANCER, PROMOTER, TFBS) should normalized, i.e. assign genomic coordinates, **if and only if the information necessary is contained in the abstract**.

It it still possible to consult external resource to determine the exact location. Nonetheless, as a rule of thumb, the whole process should take a reasonable amount of time. If the process is too lengthy (e.g. it would involve searching in the literature) the REGION should not be normalized.