UNIVERSITAT POLITÈCNICA DE VALÈNCIA

PROS
Centro de Investigación en Métodos de Producción de Software

Informe Técnico / Technical Report

Conceptual Schema of the Human Genome (CSHG)

Óscar Pastor López, José F. Reyes Román and Francisco Valverde G.

Ref. #:	**PROS-TR-2016-X**
Title:	Conceptual Schema of the Human Genome (CSHG)
Author (s):	Oscar Pastor López, José Fabián Reyes Román and Francisco Valverde G.
Corresponding author (s):	opastor@dsic.upv.es jreyes@pros.upv.es fvalverde@pros.upv.es
Document version number:	1
Final version:	-
Release date:	-
Key words:	CSHG, Conceptual models, Evolution, Human Genome, GeIS.

Tabla de Contenidos

1. Introducción

¿Por qué es esencial el modelado conceptual (CM) [1] para diseñar y desarrollar sistemas de formación adecuados? Esta es una cuestión fundamental dentro de la comunidad de CM, la cual está interesada en demostrar que sólo mediante el uso de técnicas de CM se puede lograr el diseño y desarrollo de sistemas de información de calidad.

Para ir un paso más adelante con respecto a esta cuestión, como objetivo principal de nuestro trabajo buscamos responder a esta pregunta de una manera convincente. La necesidad de una estrategia de diseño y desarrollo basado en CM debería ser más evidente mientras mayor sea la complejidad del sistema en estudio.

La comprensión del genoma humano es un buen ejemplo de un problema extremadamente complejo. El uso del CM para proporcionar una solución que haga frente al caso del genoma humano, se ha explorado inicialmente en trabajos anteriores [2-3], pero una perspectiva holística de toda la representación (imagen) todavía no se ha facilitado.

El uso de enfoques avanzados en ingeniería de SI es vital en este ámbito debido a la enorme cantidad de información biológica existente, la cual debe ser capturada, comprendida (manipulada) y controlada de manera eficaz.

Una rama importante de la bioinformática está dedicada a la gestión de los "datos genómicos", y la existencia de un gran conjunto de fuentes de datos (diversas) con grandes cantidades de datos que representan un conocimiento relacionado con la evolución continua hace que sea muy difícil encontrar soluciones convincentes.

Cuando nos enfrentamos a este problema desde una perspectiva de SI, entendimos que se precisa de la aplicación de CM para comprender la información relevante en el dominio, y así fijar con claridad y representarlo con el fin de hacer una estrategia de gestión de datos eficaz.

Sin embargo, nos sorprendió descubrir que nuestros colegas en biología no tenían idea sobre los CM de sus modelos de datos, y la respuesta recibida fue –en el mejor de los casos, simplemente un diseño lógico relacional, una descripción en el espacio de soluciones sin resumen, sin una perspectiva de diseño conceptual en absoluto.

Pronto llegamos a la conclusión de que la perspectiva de modelado conceptual era para nada utilizada, por lo que tratamos de convencer a nuestros colegas del área bioinformática para construir un Esquema Conceptual del Genoma Humano con el objetivo principal de -comprender como entendemos que funciona nuestra vida en la tierra-, permitiéndonos proporcionar una comprensión de los conceptos básicos que explican como la estructura genotípica se manifiesta en un fenotipo externo.

Nuestro objetivo es demostrar la necesidad de un CM:

- ✓ Compartir la comprensión de los conceptos esenciales del dominio -en nuestro caso el genoma humano-, y

- ✓ Orientar el diseño y el desarrollo de las correspondientes bases de datos (DB), que normalmente cubren sólo una parte de los CM. Esto significa que el uso de CM sólo como un tipo holístico, bases de datos conceptual, hará posible la integración de diferentes fuentes de datos que representan diferentes perspectivas del conocimiento genómico.

Para presentar todo el trabajo realizado siguiendo esta línea de investigación, en primer lugar, introducimos una primera representación conceptual del conocimiento genómico correspondiente, el cual denominamos Esquema Conceptual del Genoma Humano versión 1 (ECGH v1).

Después de este ejercicio conceptual, se generaron más debates en profundidad sobre cómo representar mejor los conceptos básicos donde su conocimiento asociado se mantiene en constante evolución.

Como resultado de este trabajo conceptual se presentó una extensión a la versión inicial, identificada como ECGH v1.1 hasta llegar a nuestra última versión propuesta del esquema conceptual, llamada ECGH v2.

2. Esquema Conceptual del Genoma Humano, versión 1 (ECGH v1)

La primera versión del esquema se caracteriza por ser el primer intento de abordar la descripción holística del dominio de la genómica y como tal se centra en una visión del genoma centrada en sus conceptos más básicos, obviando algunos aspectos más complejos.

El Esquema Conceptual del Genoma Humano (ECGH) versión 1 fija su atención en el análisis de genes individuales, sus mutaciones y sus aspectos fenotípicos. En consecuencia, otros fenómenos como la regulación múltiple, la codificación de una misma proteína por dos genes diferentes, los pseudo-genes o la acción combinada de múltiples genes son apartados con la intención de ser estudiadas en próximas versiones. Este primer modelo podría ser considerado como el "esencial".

Para llegar a la primera versión del ECGH se reunió a un amplio grupo de expertos en las áreas de: biología molecular y modelado conceptual, con el fin de abordar los elementos principales y esenciales del dominio de la genómica. Teniendo en cuenta de que a medida en que se adquieran nuevos conocimientos sobre el dominio se podrá generar N versiones del ECGH.

En los procesos de modelado se presentaron cuatro (4) iteraciones para llegar a la versión 1 del ECGH que podemos revisar en los trabajos [7-8].

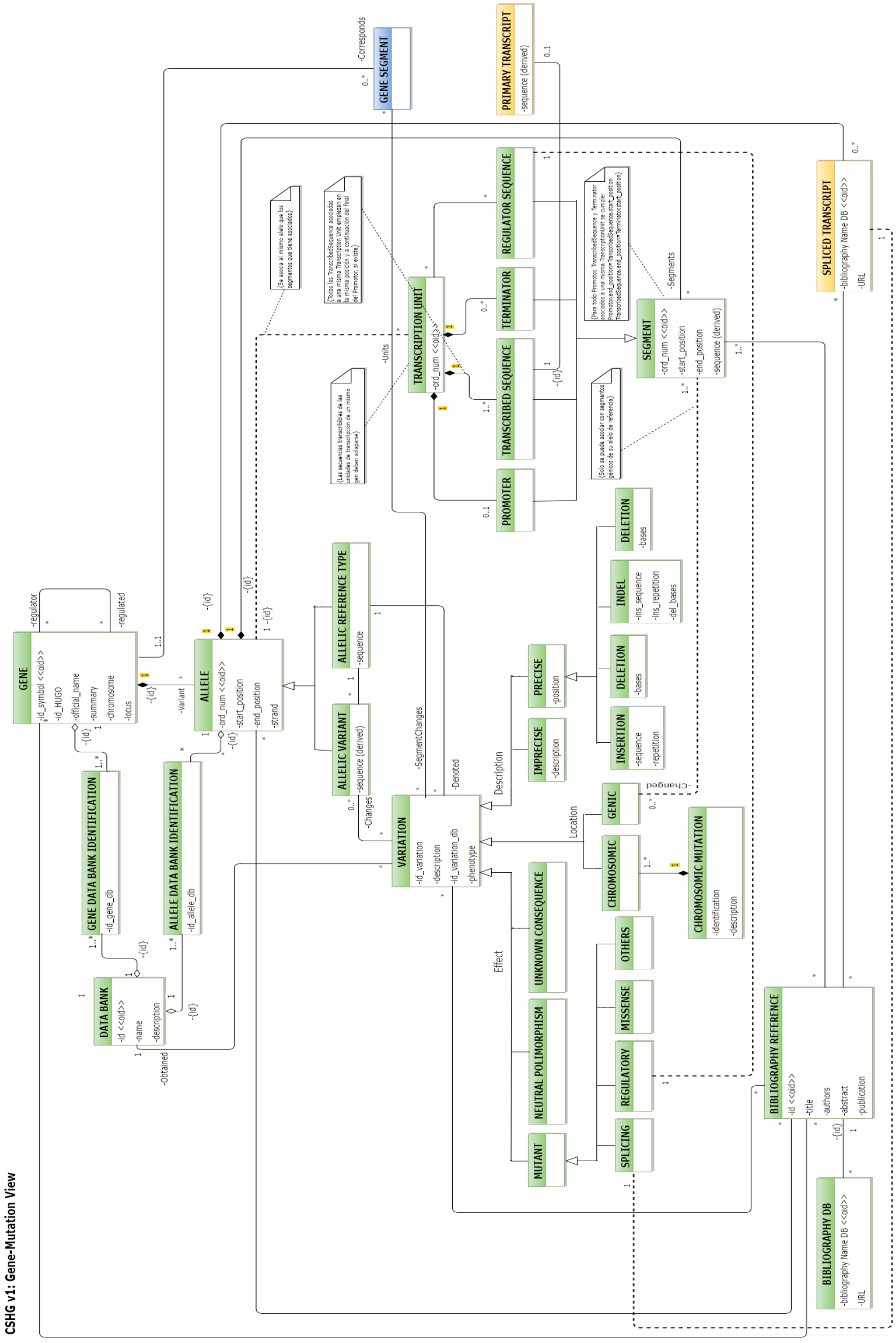
A continuación, presentamos su clasificación en tres vistas:

- **Genome View:** Esta vista se encarga de modelar genomas humanos individuales. Es una vista vital para la futura extensión del modelo. Actualmente, se trabaja en la secuenciación del genoma completo en costos más asequibles.

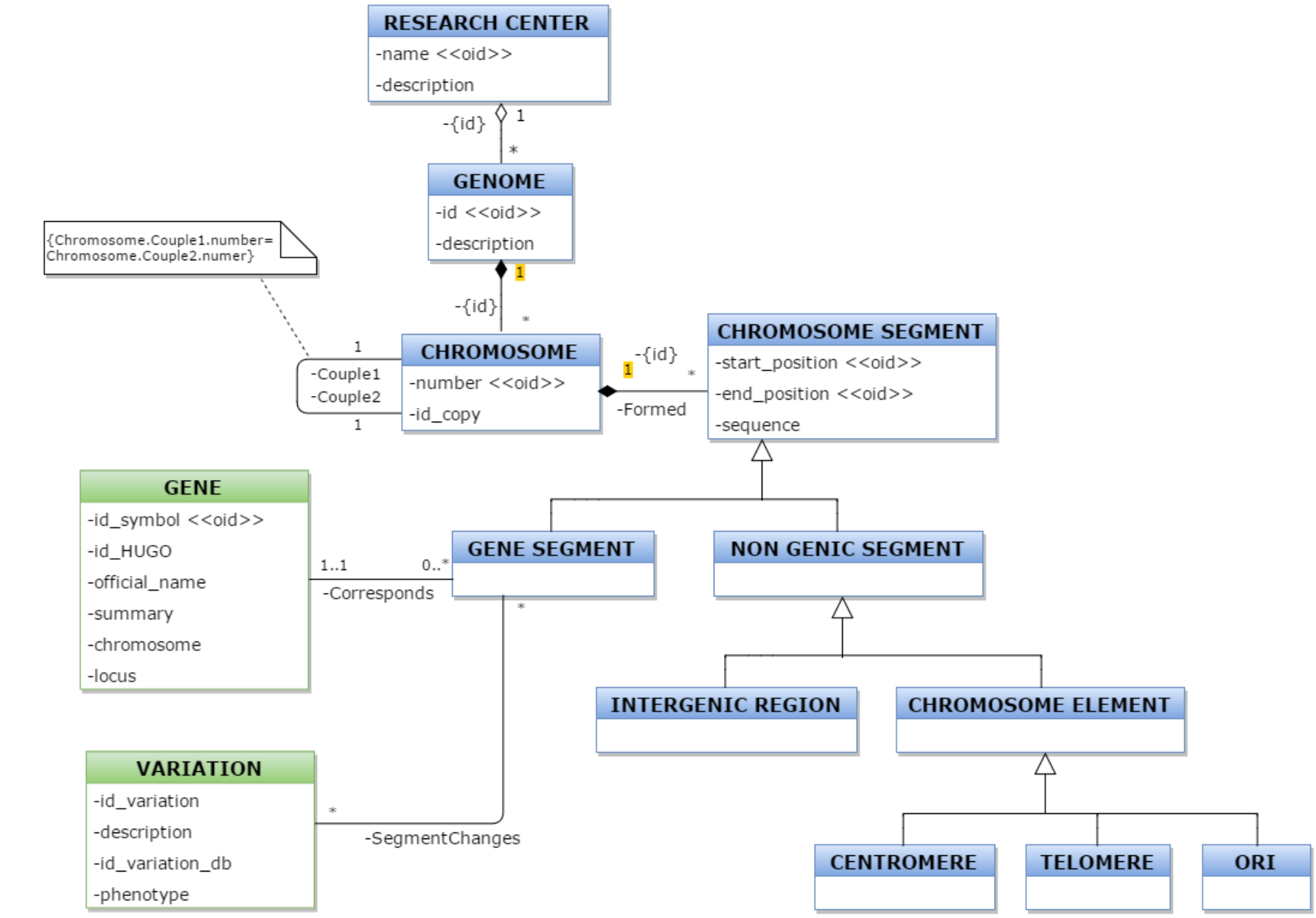
- **Gene-Mutation View:** Utilizada para modelar el conocimiento sobre los genes, su estructura y sus variantes alélicas. Las entidades con mayor relevancia que han sido modeladas en esta vista son: Gene y Allele.

- **Transcription View:** Esta vista modela los componentes básicos del proceso de transcripción y las síntesis de las proteínas (que es lo que conocemos como "expresión génica").

2.1 Gene-Mutation View

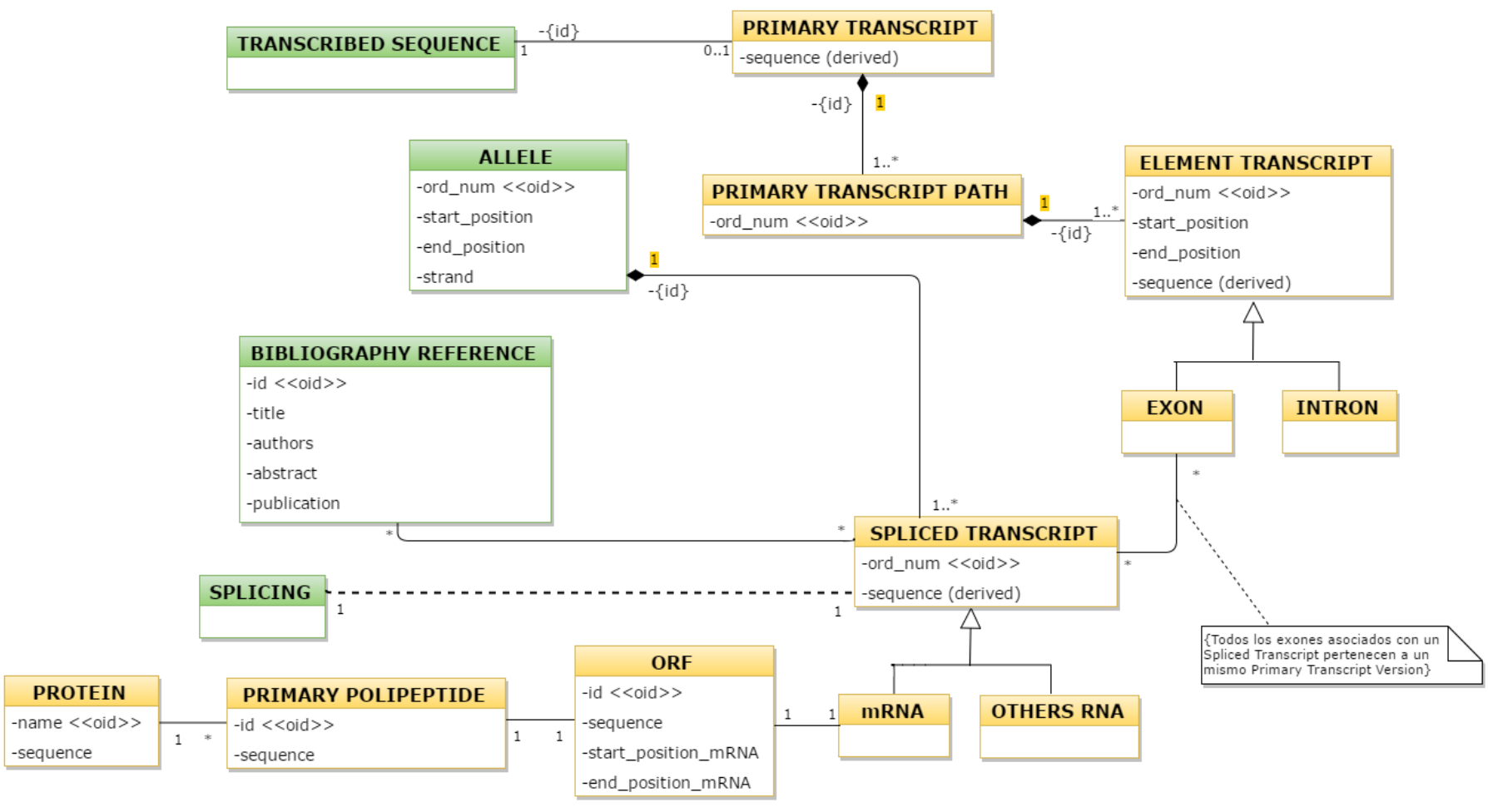


CSHG v1: Genome View



2.2 Genome View

CSHG v1: Transcription View

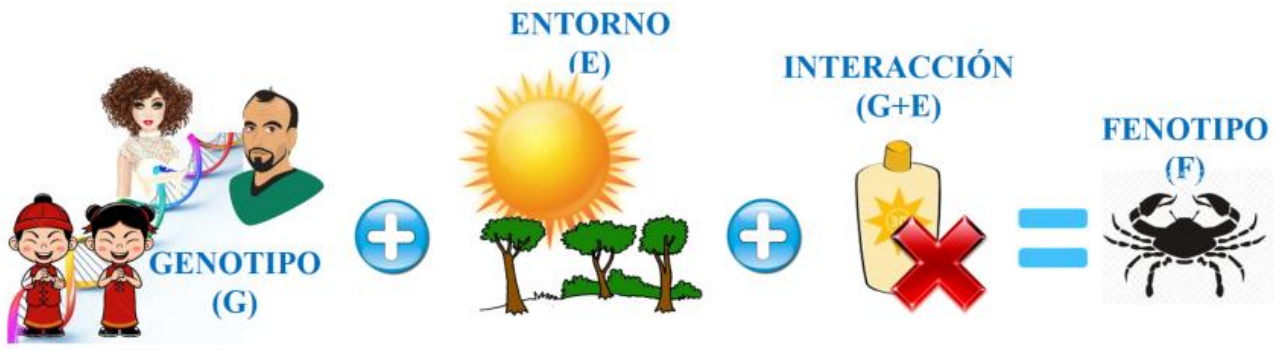


2.3 Transcription View

3. Esquema Conceptual del Genoma Humano, versión 1.1 (ECGH v1.1)

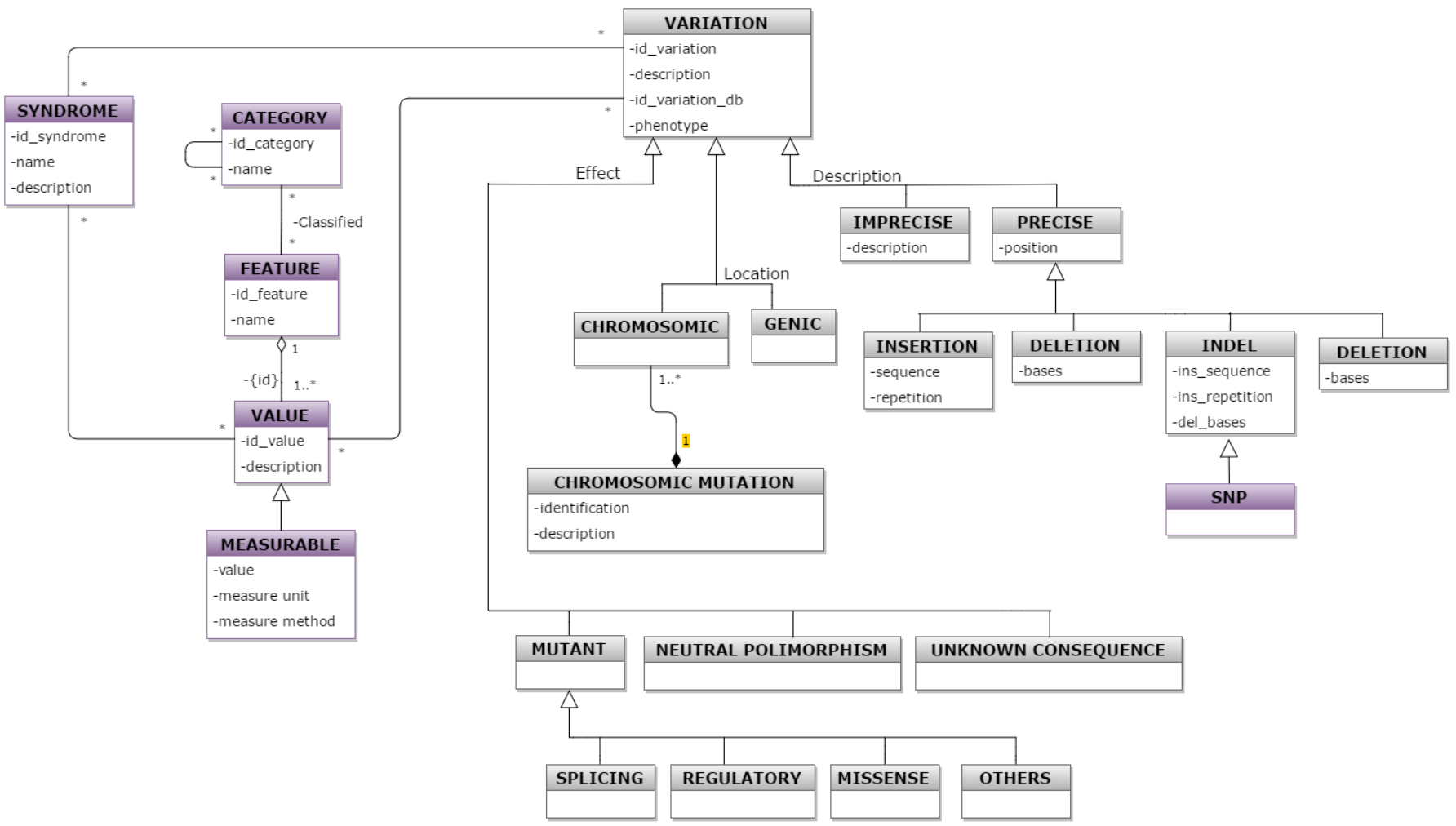
El Esquema Conceptual del Genoma Humano versión 1.1, es la evolución natural del ECGH v1 (inclusión de la vista fenotípica).

La visión fenotípica es muy importante ya que aporta mayor consistencia al esquema. El hecho de ofrecer una visión genotípica, información genética que posee un organismo, ligada a una visión fenotípica, expresión del genotipo en función de un determinado ambiente, ofrece un gran valor investigativo y dota de mayor importancia al modelo.



3.1 Genotipo y Fenotipo

Algunas nociones sobre los conceptos que fueron descritos y modelados en v1 han sufrido modificaciones debido a la continua evolución del dominio y los nuevos conocimientos que se tienen del mismo.



3.2 Phenotype View

4. Esquema Conceptual del Genoma Humano, versión 2 (ECGH v2)

Nuestra versión 2 del ECGH, cambia su núcleo central y pasa de tener una visión "gencentrísta" a una visión centrada en el concepto de cromosoma. Siendo dicho cambio de visión la principal diferencia con respecto a la versión anterior del esquema.

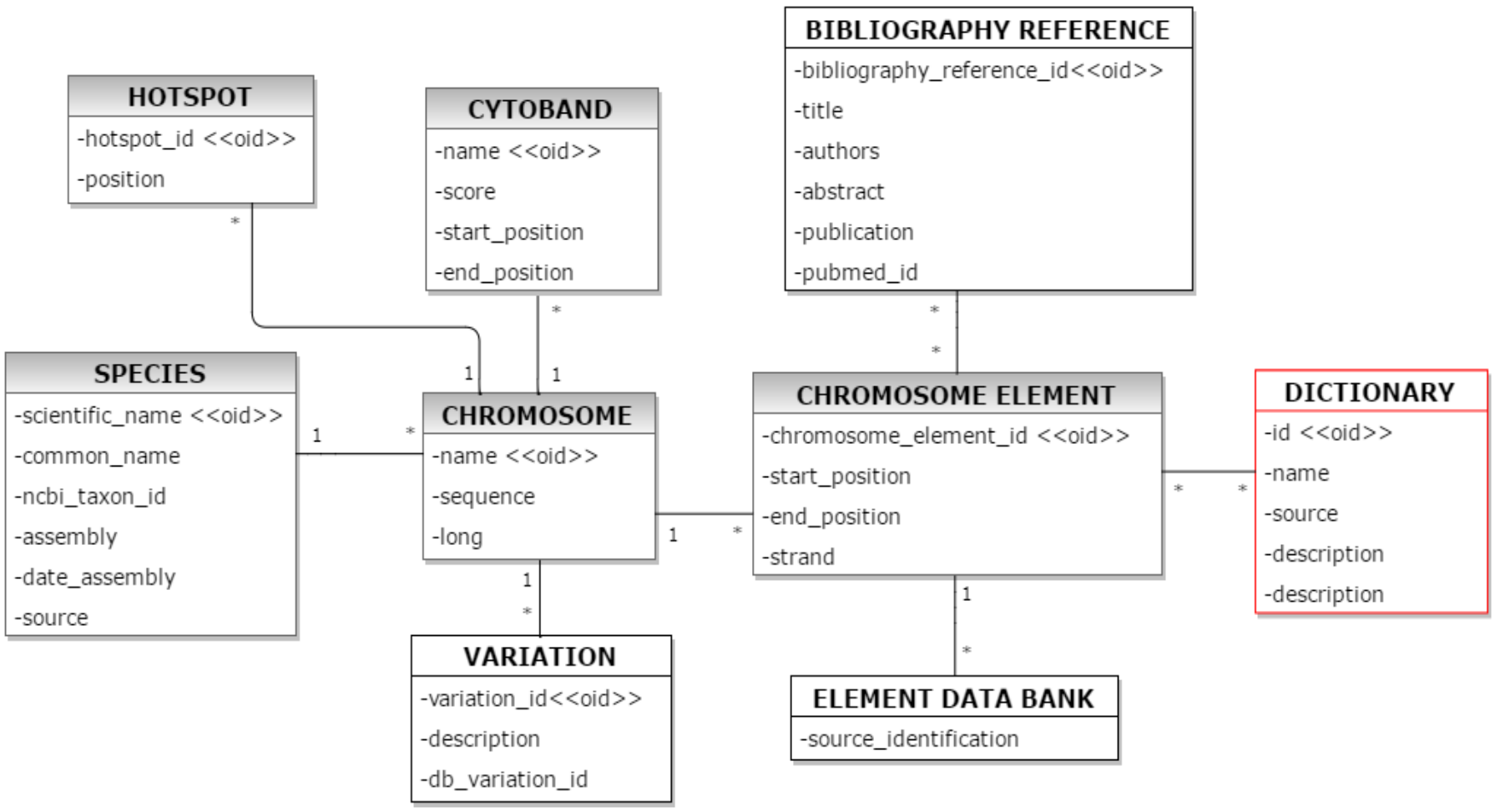
El hecho de ofrecer una visión que toma como eje central la idea de cromosoma implica diversos cambios con respecto a las versiones anteriores.

Por ejemplo, las variaciones que antes estaban referenciadas a un alelo de referencia y cuyas posiciones se definían en dicho alelo, ahora pasan a obtener sus coordenadas con respecto a un cromosoma determinado.

Además, se incorpora se extiende el modelo con nueva información, añadiendo nuevas vistas como la de *pathways*, y completando las ya existentes. A continuación, presentamos las cinco vistas que componen esta nueva versión:

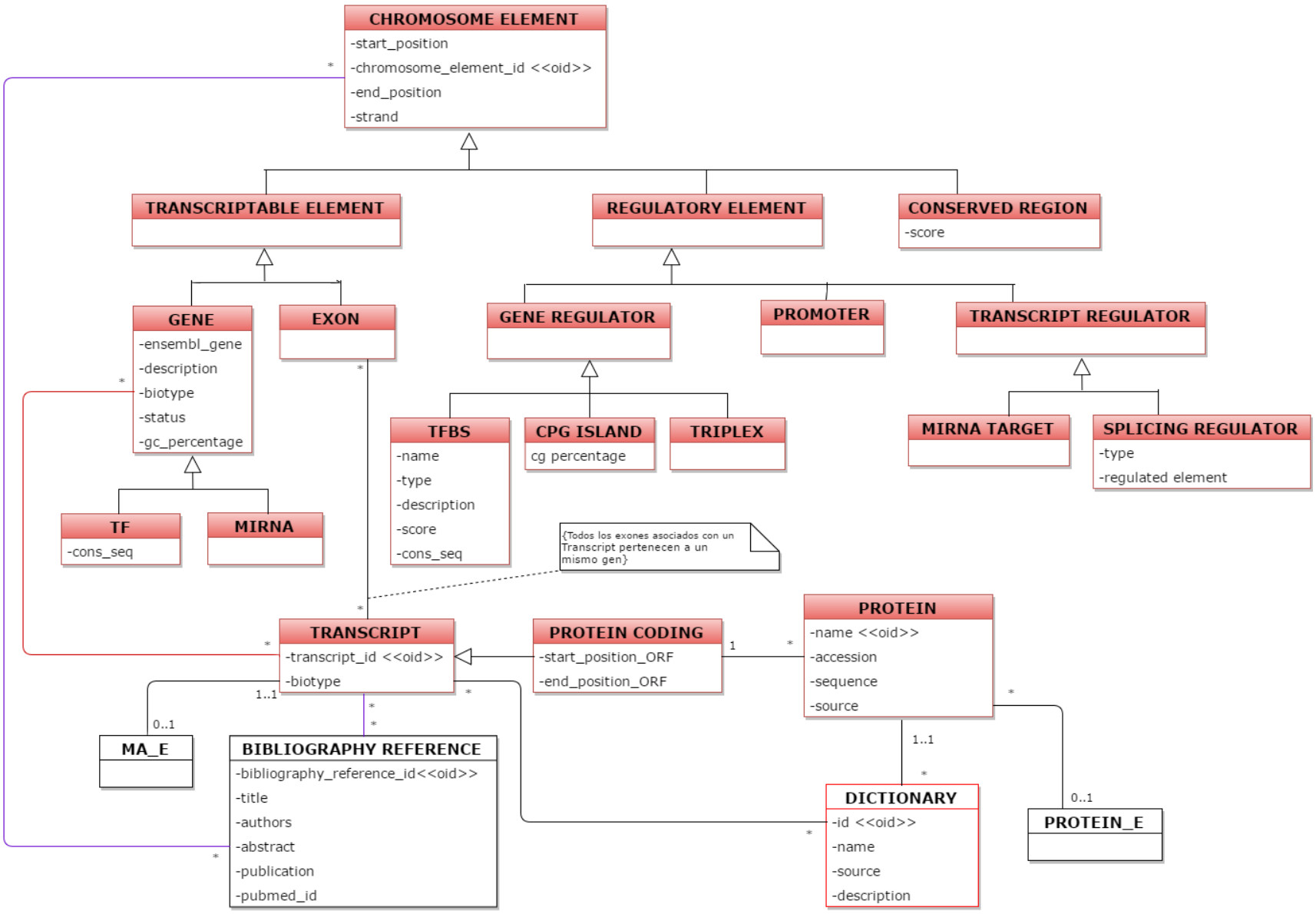
- **Vista Estructural:** Esta vista, como su nombre indica, describe la estructura del genoma. A grandes rasgos se puede decir que la información genómica en un organismo se distribuye en 23 pares de cromosomas y genes que codifican proteínas, secuencias reguladoras, etc.

  Por otro lado, cabe destacar que cada cromosoma pertenece a una única especie y que además contempla zonas calientes o hotspots y subregiones llamadas citobandas que se hacen visibles microscópicamente después del tintado.



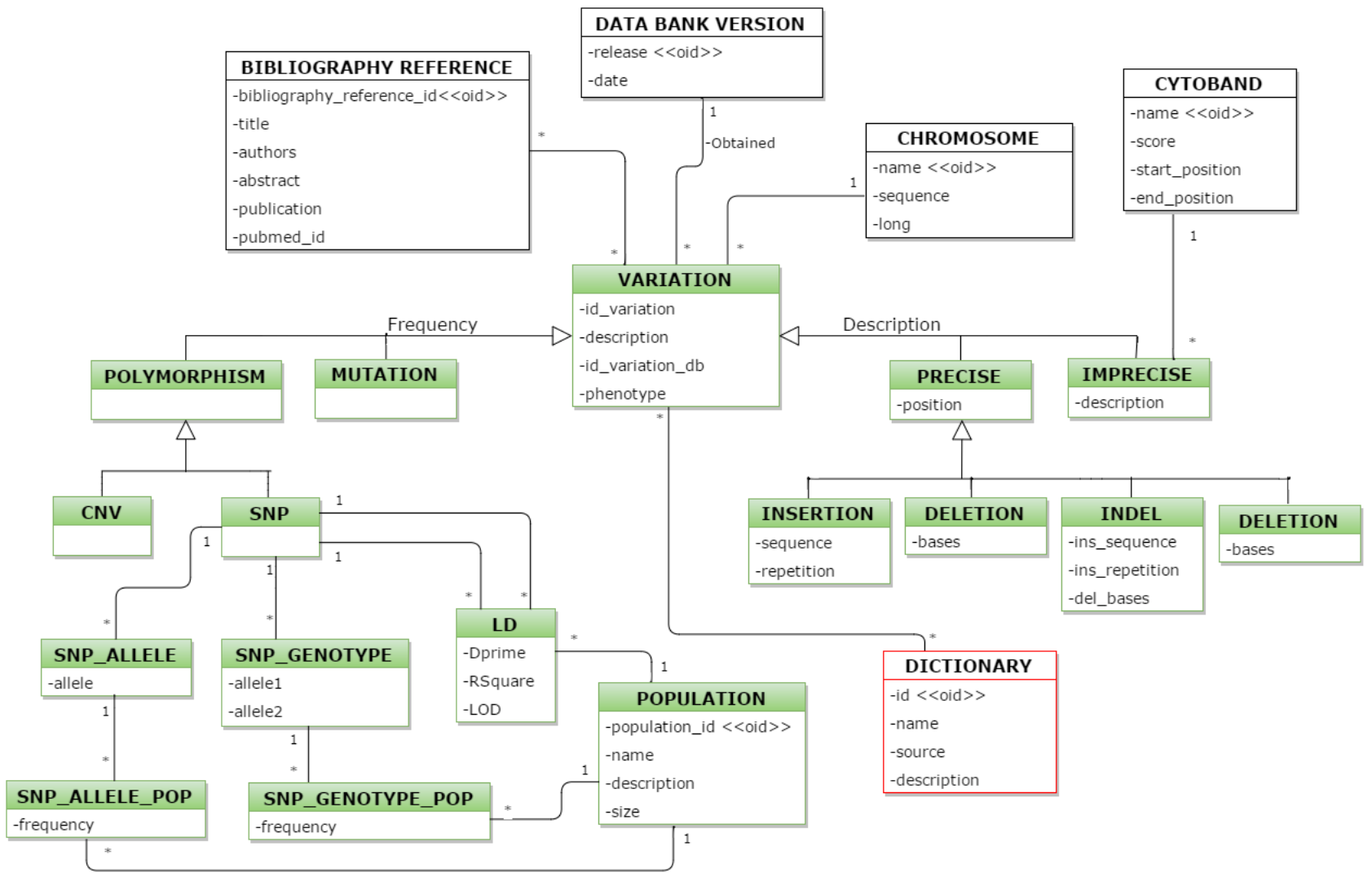
4.1 Structural View

- **Vista de transcripción:** Un gran número de genes expresan su funcionalidad a través de la producción de proteínas. La vista transcripción muestra los componentes y conceptos relacionados con la síntesis de proteínas. La secuencia de ADN que se transcribe en una molécula de ARN codifica al menos un gen, y si el gen transcrito codifica para una proteína, el resultado de la transcripción es ARN mensajero (mRNA), el cual será entonces usado para crear esa proteína a través de un proceso de traducción. Después de la transcripción, tiene lugar una modificación en el ARN llamada splicing, en la cual los intrones son borrados y los exones se unen. Pero en muchos de los casos, el proceso de splicing no es "perfecto" y puede variar la composición de los exones del mismo ARN mensajero. Este fenómeno es entonces llamado splicing alternativo. El splicing alternativo puede ocurrir de muchas maneras. Los exones pueden ser extendidos o saltados, o los intrones pueden ser retenidos.

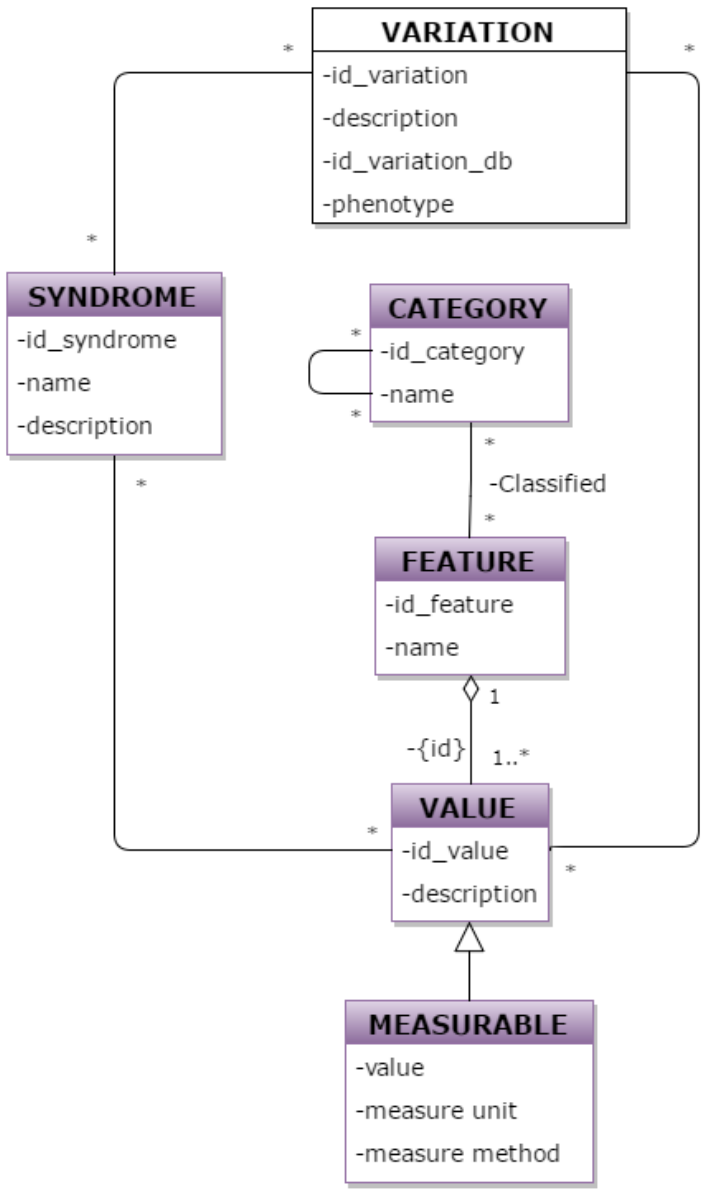


4.2 Transcription View

- **Vista de variaciones:** La vista variación modela el conocimiento relacionado con las diferencias encontradas en la secuencia de ADN de diversos individuos.



4.3 Variation View

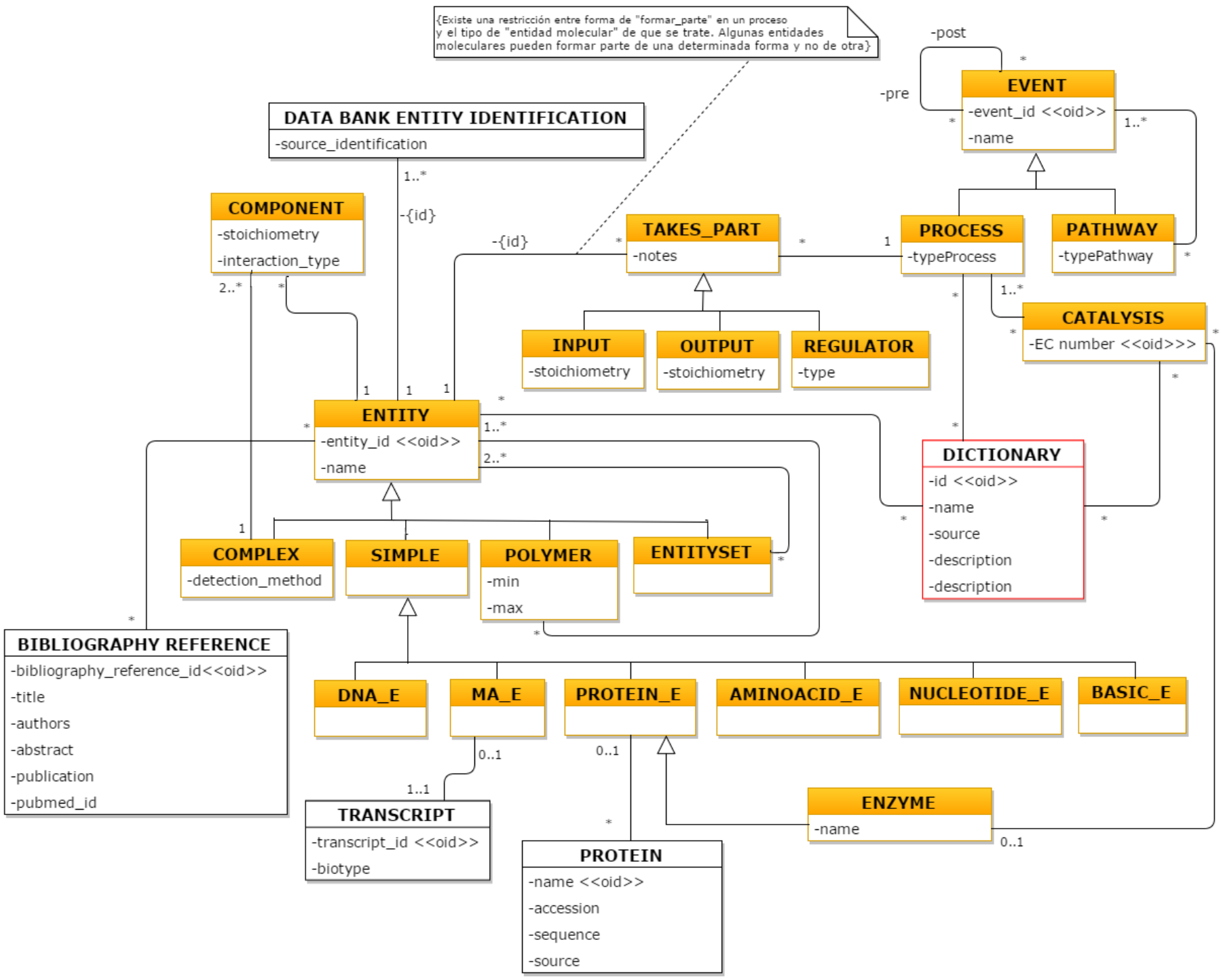


4.4 Phenotype View

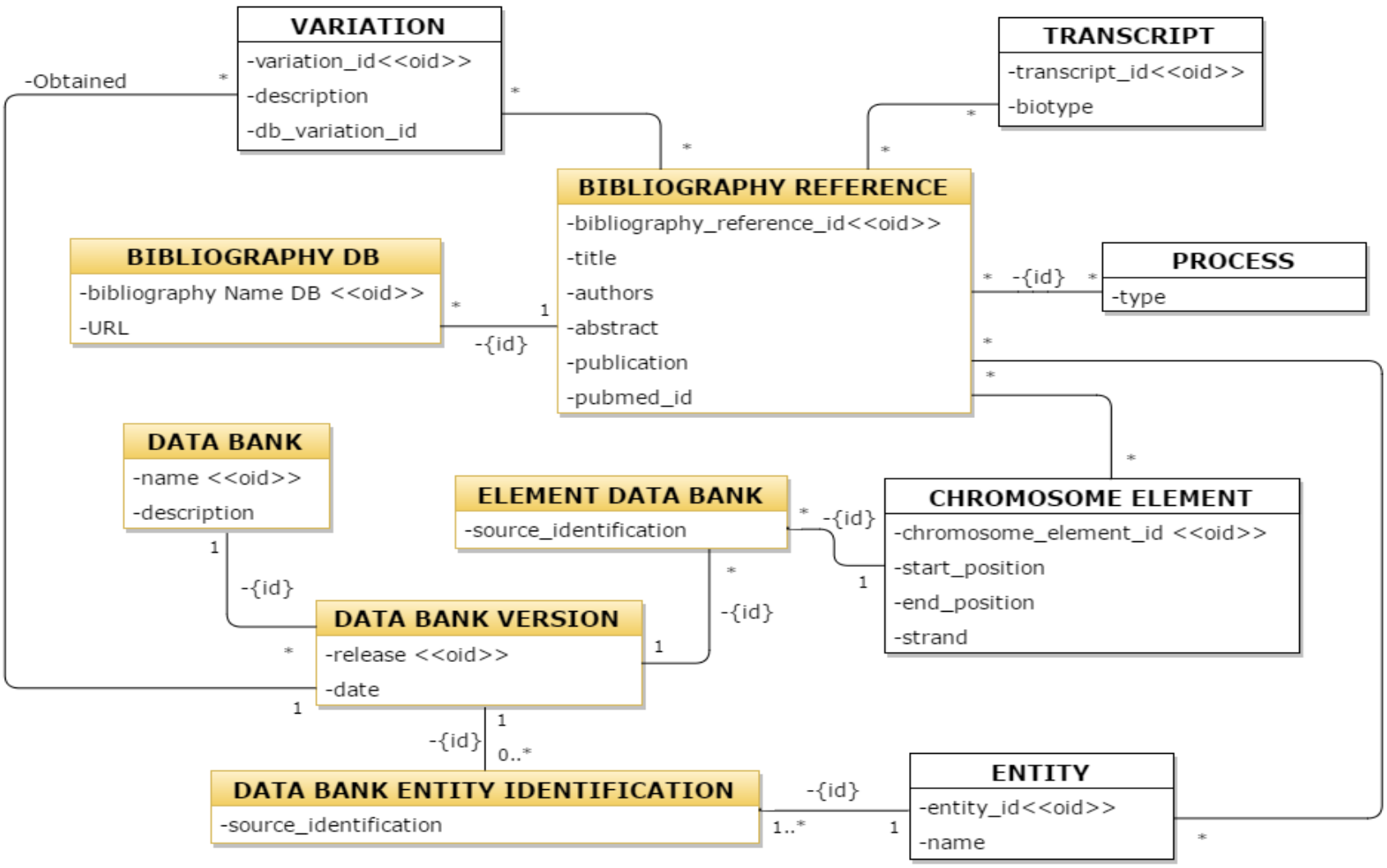
- **Vista fenotípica:** descrita anteriormente en la sección 3.

- **Vista de rutas metabólicas:** En bioquímica, las rutas metabólicas (pathways), son una serie de reacciones químicas que ocurren dentro de una célula.

- **Vista de fuentes de datos y bibliografía**: Esta vista proporciona información sobre las fuentes de datos de las que se ha extraído la información que se va a almacenar en el modelo, así como una serie de documentos bibliográficos de consulta para quien desee obtener más información con respecto a algún aspecto aquí definido. Para mantener información sobre las fuentes de las cuales se ha obtenido la información.



4.5 Pathway View



4.6 Bibliography and data bank View

5. Referencias

1. Olivé, A.: Conceptual modeling of information systems. Springer Science & Business Media (2007)
2. Bornberg-Bauer, E., & Paton, N. W.: Conceptual data modelling for bioinformatics. Briefings in Bioinformatics. Vol 3. No 2. 166–180 (2002)
3. Ram, S., & Wei, W.: Modeling the semantics of 3D protein structures. In Conceptual Modeling–ER 2004 (pp. 696-708). Springer Berlin Heidelberg (2004)
4. Rodden R., T.: Genetics for Dummies, 2nd Edition. Wiley Publishing, Inc., Indianapolis, Indiana (2010)
5. National Center for Biotechnology Information (2016). http://www.ncbi.nlm.nih.gov
6. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K.: dbSNP: the NCBI database of genetic variation. Nucleic acids research, 29(1), 308-311 (2001)
7. Pastor, O., et al.: Model driven-based engineering applied to the interpretation of the human genome. The Evolution of Conceptual Modeling. Springer, Heidelberg (2010)
8. Pastor, Oscar, et al.: Model-based engineering applied to the interpretation of the human genome. The evolution of conceptual modeling. Springer Berlin Heidelberg, 306-330 (2011)

6. Anexo: ECGH (versiones completas)

Gene-Mutation View

Genome View

Transcription View

Gene
-id_symbol <<oid>> : string
-id_HUGO : int
-official_name : string
-summary : string
-chromosome : short
-locus : string

Gene Data Bank Identificacion
-id_gene_db : string

Data Bank
-id <<oid>> : string
-name : string
-description : string

Allele
-ord_num <<oid>> : short
-start_position : long
-end_position : long
-strand : string

Allele Data Bank Identification
-id_allele_db : string

Allelic Variant
-sequence (derived) : string

Allelic Reference Type
-sequence : string

Variation
-id_variation : int
-description : string
-id_variation_db : string
-phenotype : string

Mutant
Neutral Polimorphism
Unknown consequence
Chromosomic
Genic
Splicing
Regulatory
Missense
Others

Chromosomic Mutation
-identificacion?
-description : string

Imprecise
-description : string

Precise
-position : int

Insertion
-sequence : string
-repetition : int

Deletion
-bases : int

Indel
-ins_sequence : string
-ins_repetition : int
-del_bases : int

Inversion
-bases : int

{Se asocia al mismo alelo que los segmentos que tiene asociados}

{Las secuencias transcribibles de las unidades de transcripción de un mismo gen deben solaparse}

{Todas las TranscribedSequence asociadas a una misma Transcription Unit empiezan en la misma posicio y a continuación del final del Promotor, si existe}

TranscriptionUnit
-ord_num <<oid>> : short

Promoter
TrasnscribedSequence
Terminator
RegulatorSequence

{Solo se puede asociar con segmentos génico de su alelo de referencia}

Segment
-ord_num <<oid>> : short
-start_position : long
-end_position : long
-sequence (derived) : string

{Para todo Promotor, TranscribedSequence y Terminator asociados a una misma TranscriptionUnit se cumple
Promotor.end_position=TranscribedSequence.start_position
TranscribedSequence.end_position=Terminator.start_position}

Bibliography DB
-Bibliography Name DB <<oid>>
-URL

Bibliography Reference
-id <<oid>>
-title
-authors
-abstract
-publication

Research Centre
-name <<oid>> : string
-description : string

Genome
-id <<oid>> : short
-description : string

{Chromosome.Couple1.number = Chromosome.Couple2.number}

Chromosome
-number <<oid>> : short
-id_copy <<oid>> : short

ChromosomeSegment
-start_position <<oid>> : long
-end_postition <<oid>> : long
-sequence : string

Gene-Segment
NonGenicSegment
IntergenicRegion
ChromosomaElement
Centromere
Telomere
ORI

PrimaryTranscript
-sequence (derived)

Primary Transcript Path
-ord_num <<oid>>

ElementTranscript
-ord_num <<oid>>
-start_position
-end_position
-sequence (derived) : string

Exon
Intron

Spliced Transcript
-ord_num <<oid>>
-sequence (derived) : string

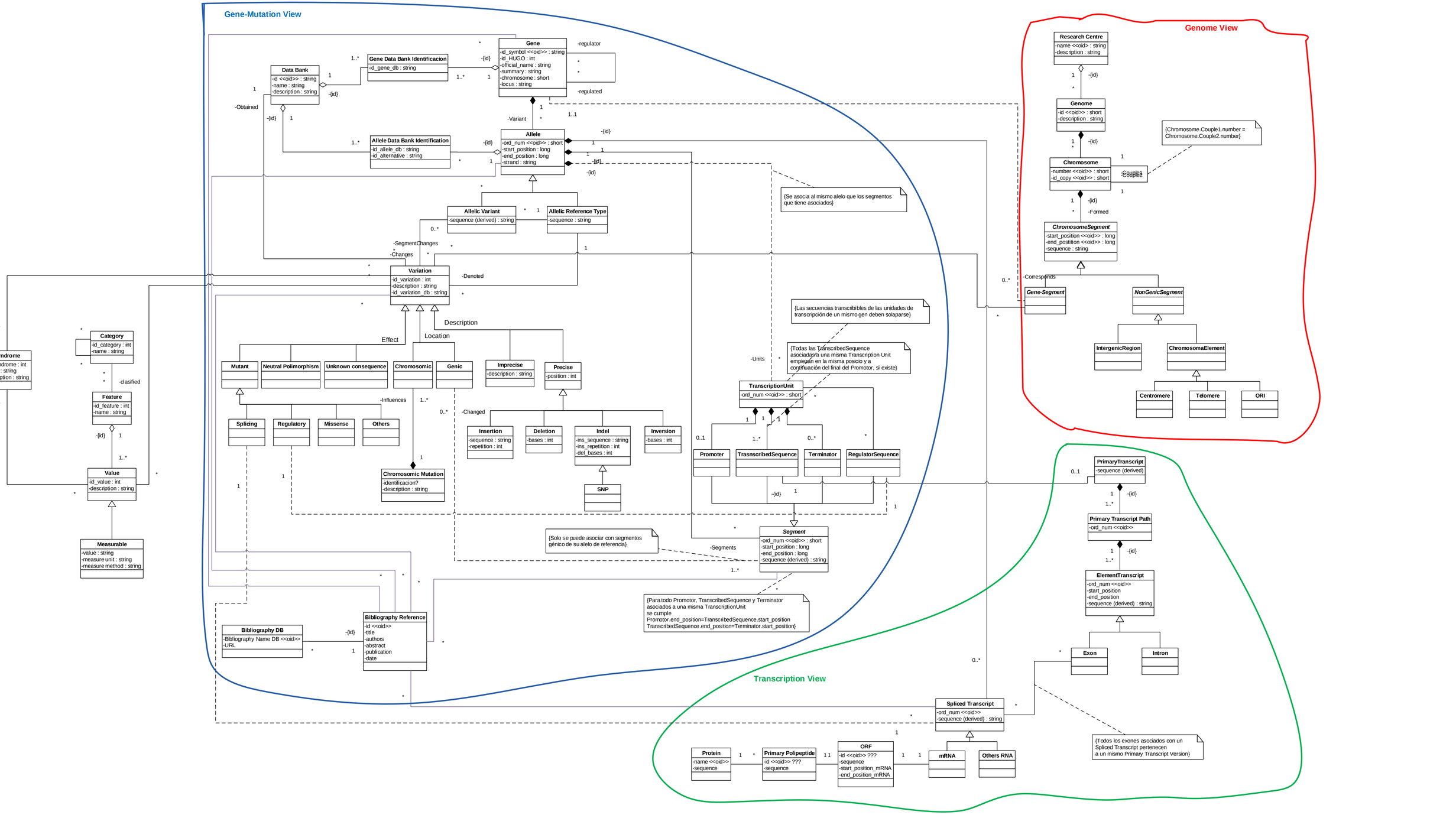
{Todos los exones asociados con un Spliced Transcript pertenecen a un mismo Primary Transcript Version}

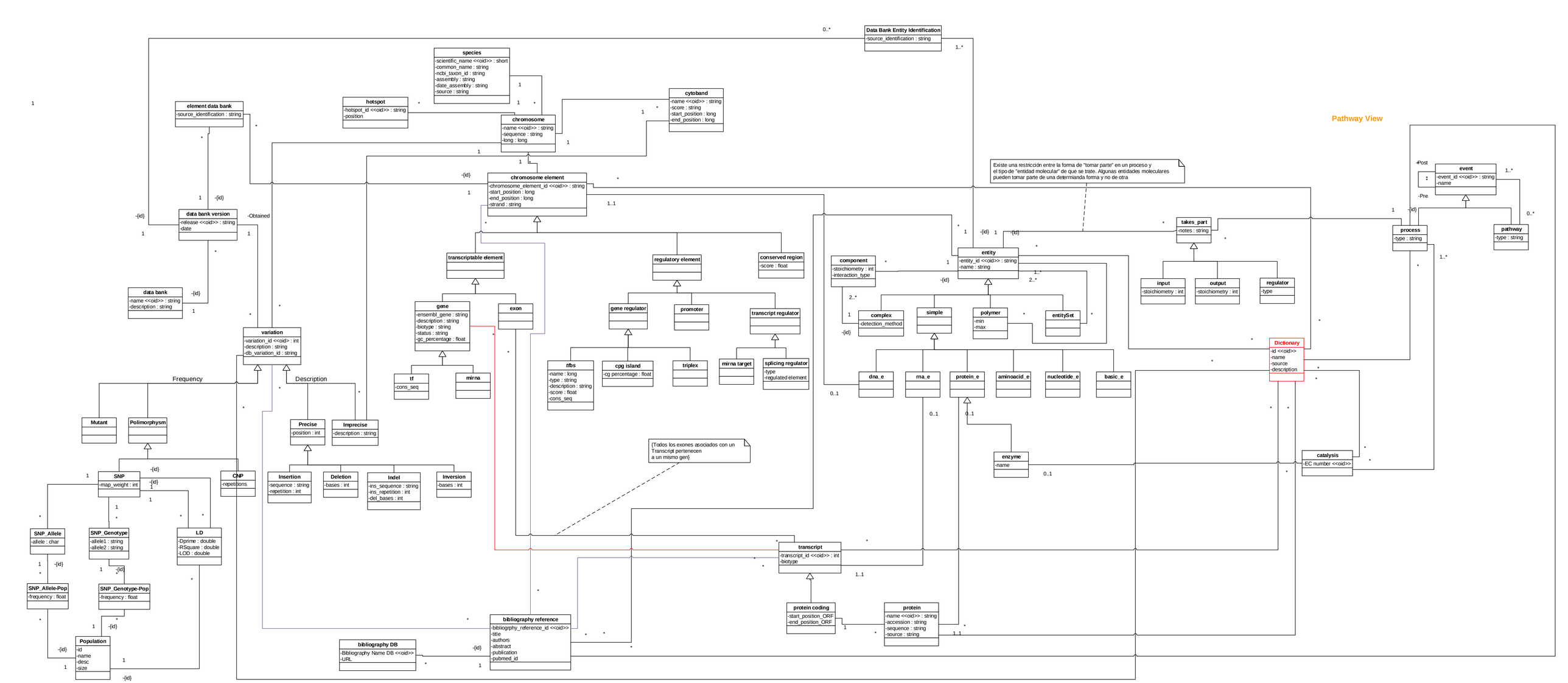
Protein
-name <<oid>>
-sequence

Primary Polipeptide
-id <<oid>> ???
-sequence

ORF
-id <<oid>> ???
-sequence
-start_position_mRNA
-end_position_mRNA

mRNA
Others RNA


Gene-Mutation View
Genome View
Transcription View
Gene
-id_symbol <<oid>> : string
-id_HUGO : int
-official_name : string
-summary : string
-chromosome : short
-locus : string
-regulator
-regulated
Data Bank
-id <<oid>> : string
-name : string
-description : string
Gene Data Bank Identificacion
-id_gene_db : string
-Obtained
-Variant
Allele
-ord_num <<oid>> : short
-start_position : long
-end_position : long
-strand : string
Allele Data Bank Identification
-id_allele_db : string
-id_alternative : string
{Se asocia al mismo alelo que los segmentos que tiene asociados}
Allelic Variant
-sequence (derived) : string
Allelic Reference Type
-sequence : string
-SegmentChanges
-Changes
-Denoted
Variation
-id_variation : int
-description : string
-id_variation_db : string
Description
Effect
Location
Mutant
Neutral Polimorphism
Unknown consequence
Chromosomic
Genic
Imprecise
-description : string
Precise
-position : int
-Influences
Splicing
Regulatory
Missense
Others
-Changed
Insertion
-sequence : string
-repetition : int
Deletion
-bases : int
Indel
-ins_sequence : string
-ins_repetition : int
-del_bases : int
Inversion
-bases : int
Chromosomic Mutation
-identificacion?
-description : string
SNP
Syndrome
-id_syndrome : int
-name : string
-description : string
Category
-id_category : int
-name : string
-clasified
Feature
-id_feature : int
-name : string
Value
-id_value : int
-description : string
Measurable
-value : string
-measure unit : string
-measure method : string
{Las secuencias transcribibles de las unidades de transcripción de un mismo gen deben solaparse}
{Todas las TranscribedSequence asociadas a una misma Transcription Unit empiezan en la misma posicio y a continuación del final del Promotor, si existe}
-Units
TranscriptionUnit
-ord_num <<oid>> : short
Promoter
TrasnscribedSequence
Terminator
RegulatorSequence
{Solo se puede asociar con segmentos génico de su alelo de referencia}
-Segments
Segment
-ord_num <<oid>> : short
-start_position : long
-end_position : long
-sequence (derived) : string
{Para todo Promotor, TranscribedSequence y Terminator asociados a una misma TranscriptionUnit
se cumple
Promotor.end_position=TranscribedSequence.start_position
TranscribedSequence.end_position=Terminator.start_position}
Bibliography DB
-Bibliography Name DB <<oid>>
-URL
Bibliography Reference
-id <<oid>>
-title
-authors
-abstract
-publication
-date
Research Centre
-name <<oid> : string
-description : string
Genome
-id <<oid>> : short
-description : string
{Chromosome.Couple1.number = Chromosome.Couple2.number}
Chromosome
-number <<oid>> : short
-id_copy <<oid>> : short
-Couple1
-Couple2
-Formed
ChromosomeSegment
-start_position <<oid>> : long
-end_postition <<oid>> : long
-sequence : string
-Corresponds
Gene-Segment
NonGenicSegment
IntergenicRegion
ChromosomaElement
Centromere
Telomere
ORI
PrimaryTranscript
-sequence (derived)
Primary Transcript Path
-ord_num <<oid>>
ElementTranscript
-ord_num <<oid>>
-start_position
-end_position
-sequence (derived) : string
Exon
Intron
Spliced Transcript
-ord_num <<oid>>
-sequence (derived) : string
{Todos los exones asociados con un Spliced Transcript pertenecen a un mismo Primary Transcript Version}
Protein
-name <<oid>>
-sequence
Primary Polipeptide
-id <<oid>> ???
-sequence
ORF
-id <<oid>> ???
-sequence
-start_position_mRNA
-end_position_mRNA
mRNA
Others RNA


Pathway View
Data Bank Entity Identification
-source_identification : string
species
-scientific_name <<oid>> : short
-common_name : string
-ncbi_taxon_id : string
-assembly : string
-date_assembly : string
-source : string
hotspot
-hotspot_id <<oid>> : string
-position
cytoband
-name <<oid>> : string
-score : string
-start_position : long
-end_position : long
chromosome
-name <<oid>> : string
-sequence : string
-long : long
element data bank
-source_identification : string
data bank version
-release <<oid>> : string
-date
data bank
-name <<oid>> : string
-description : string
chromosome element
-chromosome_element_id <<oid>> : string
-start_position : long
-end_position : long
-strand : string
transcriptable element
regulatory element
conserved region
-score : float
gene
-ensembl_gene : string
-description : string
-biotype : string
-status : string
-gc_percentage : float
exon
gene regulator
promoter
transcript regulator
tf
-cons_seq
mirna
tfbs
-name : long
-type : string
-description : string
-score : float
-cons_seq
cpg island
-cg percentage : float
triplex
mirna target
splicing regulator
-type
-regulated element
variation
-variation_id <<oid>> : int
-description : string
-db_variation_id : string
Frequency
Description
Mutant
Polimorphysm
Precise
-position : int
Imprecise
-description : string
Insertion
-sequence : string
-repetition : int
Deletion
-bases : int
Indel
-ins_sequence : string
-ins_repetition : int
-del_bases : int
Inversion
-bases : int
SNP
-map_weight : int
CNP
-repetitions
SNP_Allele
-allele : char
SNP_Genotype
-allele1 : string
-allele2 : string
LD
-Dprime : double
-RSquare : double
-LOD : double
SNP_Allele-Pop
-frequency : float
SNP_Genotype-Pop
-frequency : float
Population
-id
-name
-desc
-size
{Todos los exones asociados con un Transcript pertenecen a un mismo gen}
bibliography reference
-bibliogrphy_reference_id <<oid>>
-title
-authors
-abstract
-publication
-pubmed_id
bibliography DB
-Bibliography Name DB <<oid>>
-URL
transcript
-transcript_id <<oid>> : int
-biotype
protein coding
-start_position_ORF
-end_position_ORF
protein
-name <<oid>> : string
-accession : string
-sequence : string
-source : string
Existe una restricción entre la forma de "tomar parte" en un proceso y el tipo de "entidad molecular" de que se trate. Algunas entidades moleculares pueden tomar parte de una determianda forma y no de otra
component
-stoichiometry : int
-interaction_type
entity
-entity_id <<oid>> : string
-name : string
takes_part
-notes : string
input
-stoichiometry : int
output
-stoichiometry : int
regulator
-type
complex
-detection_method
simple
polymer
-min
-max
entitySet
dna_e
rna_e
protein_e
aminoacid_e
nucleotide_e
basic_e
enzyme
-name
catalysis
-EC number <<oid>>
process
-type : string
event
-event_id <<oid>> : string
-name
pathway
-type : string
Dictionary
-id <<oid>>
-name
-source
-description