

# ARTÍCULO DE REVISIÓN

## Soluciones ante el caos genómico

*Vanessa Solis C.<sup>1, 2</sup>, Lourdes Viñansaca<sup>3</sup>, Juan José Sáenz-Peña<sup>4, 5</sup>*

1. Centro de Investigación en Métodos de Producción de Software (PROS),  
Universitat Politècnica de València, Valencia, España
2. Departamento de Recursos Hídricos y Ciencias Ambientales (iDRHICA),  
Universidad de Cuenca, Cuenca, Ecuador
3. Facultad de Ciencias Médicas, Universidad de Cuenca, Cuenca, Ecuador
4. Escuela de Doctorado, Universitat Politècnica de València, Valencia
5. Dirección de Investigación, Universidad de Cuenca, Cuenca, Ecuador

### Correspondencia:

Vanessa Solis Cabrera

### Correo electrónico:

vsolis@pros.upv.es

**ORCID ID:** <http://orcid.org/0000-0001-9996-2565>

### Dirección:

Benigno Malo 14-69 y Rafael  
María Arizaga, Cuenca-Ecuador

**Código postal:** EC 010101

**Teléfono:** (593) 992665023

### Fecha de recepción:

20-04-2020

### Fecha de aceptación:

20-05-2020

### Fecha de publicación:

30-06-2020

### Membrete bibliográfico:

Solis-Cabrera, V. Viñansaca, L.  
Sáenz-Peña, JJ. Soluciones ante  
el caos genómico. Rev Médica  
Ateneo, 22. (1): 97-110

**Artículo acceso abierto.**

## RESUMEN

La investigación acerca del genoma humano ha dejado como resultado muchos datos e información desde su descubrimiento hasta la actualidad. Esta información es útil para investigadores, pero se vuelve un reto la extracción de información genómica por la gran cantidad de recursos disponibles en la web, esto se debe en gran parte a que no se dispone actualmente de herramientas que brinden soporte para automatizar las búsquedas personalizadas y que se discrimine la información repetida. Este trabajo pretende plantear y forjar soluciones que mitiguen los problemas que se tienen en el momento de la investigación genómica. Partiendo de una exploración de bases de datos genómicas orientadas al humano, donde se han podido evidenciar problemas que conllevan al conocido "caos genómico", y a raíz de ello se plantean soluciones que ayuden a controlarlo y disminuirlo. De esta manera poder integrar la información de mejor manera, filtrar la información según las necesidades de investigación, tener un manejo adecuado de los datos, extraer información sin importar la presentación de estos y finalmente preservar la información actualizada.

**Palabras clave:** Caos genómico, Análisis de información, Bases de datos genómicas, Esquema conceptual del genoma humano (ECGH), Genoma humano.

## ABSTRACT

Research on the human genome has resulted in much data and information from its discovery to the present day. This information is useful for researchers but an extraction of genomic information is retrieved due to the large number of resources available on the web, this is largely due to the fact that tools that provide support to automate personalized searches are not currently available and that repeated information is discriminated. This work aims to propose and forge solutions that mitigate the problems they have at the time of genomic research. Starting from an exploration of human-oriented genomic databases, where evidence of problems that lead to the known "genomic chaos" has been found, and one root of this is solutions that help control and reduce it. In this way, being able to integrate the information in a better way, filter the information according to the research needs, have an adequate handling of the data, extract information regardless of the presentation of the data, and finally preserve the updated information.

**Key words:** Genomic chaos, Information Analysis, Genomic databases, Conceptual schema of the human genome (CSHG), Human Genome.

## INTRODUCCIÓN

El estudio del ADN ha evolucionado constantemente de manera gradual, es así como ha cumplido grandes hitos históricos, tales como: en 1909 se lo pudo identificar químicamente, en 1953 se definió su estructura donde por primera vez se pudo observar que existían 2 hélices (1). Sin embargo, recién en 1990 se conformaron diferentes proyectos para poder decodificar y conocer los diferentes componentes del genoma, muchos institutos han recabado información que poco a poco ha sido de gran ayuda (2), y en 2001 se realiza la primera secuenciación del genoma humano (3). A lo largo de este tiempo la información genómica se ha recogido y colgado en la web a manera de recursos informáticos (4), pero lamentablemente no todos los sitios cuentan con una supervisión que avale todo lo que se encuentra publicado, adicionalmente se carece de una estandarización o normativa, por lo que la información suele ser escueta, presentada en diferentes formatos y lenguajes, es ahí donde se ha generado el llamado "caos de datos genómicos"(5) (6).

Los investigadores tienen a su disposición una amplia variedad de información genética y gran parte de esta información se encuentra en repositorios que pertenece a los grupos de investigación, así como en sitios web especializados (7). No obstante, el reto para la sociedad de investigadores es poder recabar información genómica que permita descubrir o enlazar datos de relevancia científica con nuevos hallazgos médicos, todo enmarcado en una investigación genómica validada (2). En este trabajo se trata este tema desde una perspectiva genómica e informática.

La perspectiva genómica aborda la heterogeneidad de la información y la dispersión de los datos que permiten recopilar, almacenar, procesar y distribuir la información genómica generada a gran escala y en lo posible actualizada.

La perspectiva informática aborda diversas problemáticas como la presentación de la información al momento de buscar, la nomenclatura es diferente entre el investigador y los sistemas de información, diferencias en el texto y formatos al presentar la información genética, inconvenientes al extraer información de las mutaciones, el idioma como limitante

en algunos sistemas de información, así como la falta de mantenimiento lo que los vuelve obsoletos y el acceso libre o por pago que afecta de forma indirecta.

En base al análisis de algunos sistemas de información genómica humana, se proponen soluciones para resolver las problemáticas planteadas con la intención de homogenizar los datos e integrar en recursos que manejen las diferentes fuentes de datos de manera estándar para poder preservar y usar dicha información.

### **Problemática general**

Uno de los principales problemas encontrados en el momento de realizar una búsqueda de información genómica es cuando la información presentada en las diferentes fuentes mantiene un grado de dispersión, heterogeneidad, redundancia y muchas de las veces inconsistencia, de esta manera para los encargados de la búsqueda de información genómica es complicado el tratamiento de los datos que se presentan como resultados. A esta problemática se la conoce como "Caos de datos genómicos". Es así como, el dominio del tema es fundamental (conocimiento del mundo genómico), en conjunto con el conocimiento de la estructura de datos que presenta cada uno de los sitios web correspondientes a las bases de datos (BD) orientadas al genoma humano.

Cada base de datos genómica (BDG) o repositorio presenta una estructura diferente de información, además las nomenclaturas que se manejan y utilizan en el momento de consultar la información, por lo que la persona que se encuentra recabando dicha información debe poseer conocimientos mínimos y básicos que permitan guiar y conducir a la investigación que se está realizando. Por este motivo, la recopilación y búsqueda de información resulta extenuante y en muchos casos se la debe gestionar de forma manual para no descartar datos importantes por cambio de nomenclaturas o abreviaturas, "similitudes" o "sinónimos" en terminología genética que pudiesen dar a confusión.

Como mecanismo para encontrar los diferentes problemas que conlleva el "Caos genómico" se trabajó con una base de conocimientos genérica en la que se consideró 2 aspectos: el Esquema Conceptual del Genoma Humano (ECGH) y una exploración de base de datos genómicas orientadas al humano (BDGH). El ECGH propuesto por Reyes (8) se centra en la información que simplifica el genoma humano para estudios en general, mientras que la exploración de BDGH realizada por Solís (9) evalúa las diferentes BDG públicas de acceso gratuito habilitadas. En el mapeo de información del ECGH y la exploración de BDGH visibilizó algunos problemas que conlleva la búsqueda de información genómica. Como resultado de ello a continuación se presenta los problemas encontrados que conllevan al "Caos genómico" tanto desde la perspectiva genómica como informática.

### **Análisis del Caos Genómico**

A la hora de investigar es indispensable conocer el ámbito en el que se trabaja y sobre todo analizar las fuentes de información disponibles para la temática. Es así como los investigadores al momento de acceder a la información empiezan a encontrar problemas y retos. En el caso de la búsqueda y análisis de datos genómicos se han podido encontrar problemas derivados del caos genómico como por ejemplo: sistemas de información genómica de acceso libre y otros

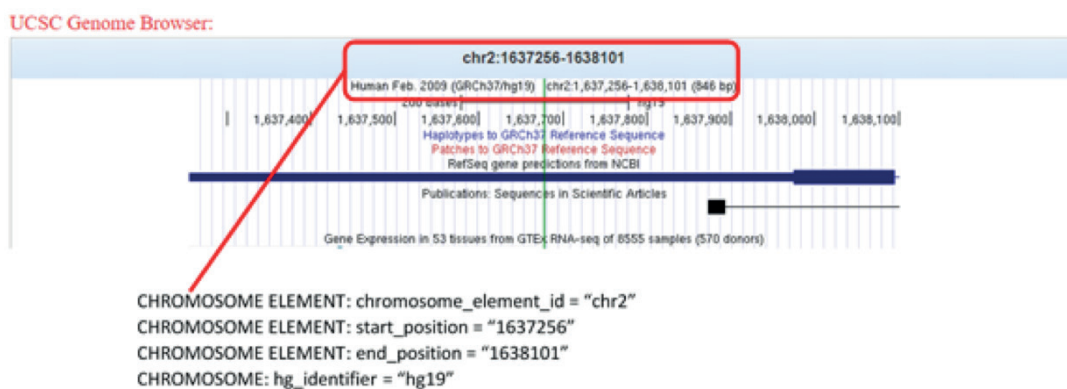
de versión profesional, la información en variedad de formatos, algunos con herramientas de búsqueda básicas y otras que tienen variedad de herramientas de búsqueda avanzadas, la información se actualiza por rangos de algunos años, otras por meses y otras se actualizan a diario, los datos deben estar curados y validados para garantizar la calidad de los datos de entrada. Por este motivo, para visibilizar el caos genómico se ha agrupado en 5 aspectos: Integración (forma en la que la información puede ser utilizada o se encuentra en la web), Filtrado de información (como se encuentra la información en los diferentes repositorios), Manejo de datos (información genómica que se tiene en cada repositorio), Presentación de datos (como se visualiza la información en las diferentes plataformas) y Mantenibilidad (actualización de la información en los diferentes repositorios).

## Integración

Una de las problemáticas dentro del ámbito genómico es la dispersión de la información de los datos suministrados por las bases de datos, repositorios en la web, software o plataformas creadas para ello, pero que no manejan los mismos lineamientos por lo que existen datos dispersos. Por este motivo, el reto es integrar, homogenizar, analizar, actualizar para que sea de fácil e inmediato acceso al seleccionar los datos en las áreas que se requiera. La información genómica del genoma humano está disponible, pero a partir de esto han surgido nuevas interrogantes y por ende investigación de las alteraciones y modificaciones que se pudieran encontrar en los 3000 millones de pares de bases aproximadamente que conforman los 50.000 genes que se encuentran en la estructura de los 46 cromosomas de la especie humana (8) (10).

La presentación de la información de las diferentes BDG no se encuentra estandarizada por lo que en algunos casos los datos se muestran en tablas otros en gráficos. Para visibilizar este problema se toma la base de datos "USCG Genome Browser" (11), en este caso la presentación de la información se la hace de forma gráfica. En la Ilustración 1 se puede observar que la localización del gen se encuentra completo, pero es complicado poder realizar la extracción de dicha información mediante un script normal, por este motivo se debe depurar de otra manera la imagen visualizada.

**Ilustración 1. Presentación de información en USCG Genomic Browser.**



En diferentes circunstancias para la persona que investiga o el software que extrae la información desde las BDG, se tiene el inconveniente al momento con la información de mutaciones, debido a que en un solo campo consta también la operación en conjunto con los elementos que intervienen, pero para manejo se suele requerir dicha información por separado. Como ejemplo a este problema se toma la base de datos genómica "Autosomal Dominant Polycystic Kidney Disease" (12). La Ilustración 2 indica el tipo de mutación correspondiente en el ECGH a "MUTATION" de manera similar se considera en base a este campo la información que presenta la columna "cDNA Change" ya que en ella se indica la operación que se realizó.

**Ilustración 2. Presentación de información para el caso de mutaciones**

**Autosomal Dominant Polycystic Kidney Disease: Mutation Database**

PKD FOUNDATION  
Polycystic Kidney Disease

Main Page

Gene: PKD1 PKD2 Mutation: Mutation Type: Clinical Significance: Region: Codon: Exon: Intron: Show All Search

Total Number Of Records Matching Criteria = 2323 2089 = Total Number Of Unique Pedigrees  
Unique pedigrees are not recorded for mutations classified as Likely Neutral

| Row | Region       | Code | Mutation Designation  | cDNA Change | Amino Acid Change | Mutation Type  | Clinical Significance | Score | # | %    |
|-----|--------------|------|-----------------------|-------------|-------------------|----------------|-----------------------|-------|---|------|
| 1   | 5'E4F1-EX15  | 1    | 5'E4F1-EX15del150k... | 1_6915del*  | Met1fa            | LARGE DELETION | Definitely Pathogenic | 1 (1) | 1 | 100  |
| 2   | 5'RAB26-EX21 | 1    | 5'RAB26-EX21del63k... | 1_3011del*  | Met1fa            | LARGE DELETION | Definitely Pathogenic | 1 (1) | 1 | 100  |
| 3   | 5'IVS1       | 1    | 5'IVS1del2.5kb        | 1_215del*   | Met1fa            | LARGE DELETION | Definitely Pathogenic | 1 (1) | 1 | 100  |
| 4   | 5'UTR        |      | -117G>T               | -117G>T     | Silent 5'UTR      | SUTR           | Likely Neutral        | - (1) | 1 | Rare |
| 5   | 5'UTR        |      | -108C>T               | -108C>T     | Silent 5'UTR      | SUTR           | Likely Neutral        | - (1) | 1 | Rare |
| 6   | 5'UTR        |      | -76G>C                | -76G>C      | Silent 5'UTR      | SUTR           | Likely Neutral        | - (1) | 1 | Rare |
| 7   | 5'UTR        |      | -67C>T                | -67C>T      | Silent 5'UTR      | SUTR           | Likely Neutral        | - (1) | 1 | Rare |
| 8   | 5'UTR        |      | -61T>C                | -61T>C      | Silent 5'UTR      | SUTR           | Likely Neutral        | - (1) | 1 | Rare |

MUTATION= Mutation Type  
SNP\_GENOTYPE: allele1 = "6915"  
SNP\_GENOTYPE: allele2 = ""

Para el problema de estandarización de información una de las soluciones propuestas es utilizar el resultado del mapeo de cada BDGH con el MCGH. De esta manera se tendrá la información mínima necesaria con los criterios y conceptos iguales. Sin embargo, en la actualidad no existe una estandarización de esta información por lo que los sitios web no están obligados a modificar su información, pero se debe plantear que a futuro dicha información pueda tener parámetros mínimos a cumplirse.

Con una exploración de la información que manejan las diferentes bases de datos genómicas públicas actuales se ha podido evidenciar que la heterogeneidad de formatos, la automatización de la exploración y extracción de información son aún ambiguas, ya que en su gran mayoría los sitios web contienen información escueta. El porcentaje de uso de servicios Web, APIs u otros mecanismos de acceso es aún bajo en BDG públicas, lo cual imposibilita el manejo de la información e integración con otros sistemas. La solución a esto podría ser la creación de scripts que usen web scraping para extracción de datos, pero que tendrían que ser personalizados para cada BDG. En el caso de herramientas de extracción de información se debe tener en cuenta las diferentes opciones de nomenclaturas que se maneje de esta manera el resultado de la extracción y filtrado de datos que se recabará para el investigador sería la más adecuada.

## Filtrado de información

Los recursos informáticos disponibles permiten filtrar los datos del genoma por cromosoma, por gen, variantes de un gen, analizar secuencias del genoma humano normal, pero se debe considerar la cantidad de mutaciones, la variabilidad de grupos de población, la variabilidad individual dentro de un mismo grupo, el movimiento migratorio humano que influyen en la creciente generación de datos y que exista gran cantidad que no son revisados, no son curados, no son validados manifestándose la dispersión de la información de datos del genoma humano. (13)

Uno de los problemas frecuentes es el texto que se utiliza en la presentación de las diferentes BDG, para ejemplificar esto se utilizará la base de datos genómica "TransmiR" (14). Como se puede ver en la Ilustración 3, el campo "binding site" se refiere a la localización (Cromosoma, inicio, fin), el color verde la interpretación de "Specie" es representado en el ECGH como "scientific\_name" pero en la mayoría de las bases de datos el nombre científico es Homo Sapiens, en este caso se lo encuentra abreviado como "H. sapiens".

### Ilustración 3. Presentación de información indicando la terminología diferente que se utiliza, en este caso para ejemplificar se tiene la base de datos genómica TransmiR

You can search the entries by such keywords:

H.sapiens miRNA exact hsa-mir-200a TF Example miRNA Example

Click to Search Reset all

| TF name | miRNA name   | TSS           | Binding site                     | Action type | SRAID/PMID | Evidence   | Tissue          | Species   |
|---------|--------------|---------------|----------------------------------|-------------|------------|------------|-----------------|-----------|
| AKT2    | hsa-mir-200a | n/a           | n/a                              | Regulation  | 22809628   | literature | n/a             | H.sapiens |
| AR      | hsa-mir-200a | chr1: 1167104 | chr1: 1164686-1164783(score=350) | Regulation  | SRX250092  | level 1    | Breast          | H.sapiens |
| AR      | hsa-mir-200a | chr1: 1167104 | chr1: 1166120-1166340(score=953) | Regulation  | SRX433201  | level 1    | Prostate        | H.sapiens |
| ARNTL   | hsa-mir-200a | chr1: 1167104 | chr1: 1162806-1163020(score=577) | Regulation  | SRX666557  | level 1    | Breast          | H.sapiens |
| ASCL2   | hsa-mir-200a | n/a           | n/a                              | Repression  | 25371200   | literature | n/a             | H.sapiens |
| ATF2    | hsa-mir-200a | chr1: 1167104 | chr1: 1162787-1163152(score=659) | Regulation  | SRX359880  | level 1    | Digestive tract | H.sapiens |
| BMP4    | hsa-mir-200a | n/a           | n/a                              | Activation  | 20621051   | literature | n/a             | H.sapiens |
| CBX3    | hsa-mir-200a | chr1: 1167104 | chr1: 1162863-1163021(score=450) | Regulation  | SRX190214  | level 1    | Digestive tract | H.sapiens |

CHROMOSOME ELEMENT: chromosome\_element\_id = "chr1"  
CHROMOSOME ELEMENT: start\_position = "1164686"  
CHROMOSOME ELEMENT: end\_position = "1164783"

SPECIE: scientific\_name

## Manejo de datos

Los sistemas de información almacenan a gran escala información genómica, por lo que es necesario comprender el funcionamiento de estos sistemas para una mejor comprensión y discriminación de los datos. Los Sistemas de información genómica (GeIS, Genimic Information Systems) permiten recopilar, almacenar, procesar y distribuir la información genómica que está en constante actualización alimentando la base de datos, adicionando o dejando de lado recursos obsoletos; para que esté a disposición de los profesionales que requieren de estos datos y que sea altamente eficientes al disponer de gran cantidad de información en una sola base de datos (8).

Los sistemas de información se presentan en bases de datos que pueden ser amplias considerando la variabilidad genética, genética de poblaciones; también pueden ser

delimitadas para genes específicos o determinada patología como el cáncer, así como disponer de herramientas para consulta de información genómica relacionada que permite conocer protocolos experimentales, diseño de secuencias normales y anormales que desde el punto de vista diagnóstico permite reproducir un experimento, diseñar cebadores, seleccionar reactivos, etc.

## Presentación de datos

Estos problemas han surgido porque la información que se presenta en los diferentes recursos genómicos, algunos de ellos siendo bases de datos genómicas puede exponerse de diferentes maneras, es decir formatos, presentación (texto o gráficos), idiomas, estructura de datos, entre otros.

El conocimiento básico del tema genómico es primordial a la hora de realizar una investigación para poder diferenciar los datos que son presentados en las BDG. En base a los atributos que han sido presentados en el ECGH se tiene una idea de la información básica sobre el genoma humano. Para ejemplificar este problema consideramos la base de datos genómica BioGPS (15) en donde se detecta que en la tabla de presentación de información el campo "Genome location" pertenece en el ECGH a diferentes atributos como se aprecia en la Ilustración 4 enmarcado en color verde, así mismo en color naranja "Aliases" que se asocia con "gene\_synonym" dentro del ECGH (9).

**Ilustración 4. Forma de presentación d la información**

|                                                                                                                       |                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|-----------------------------------------------------------------------------------------------------------------------|--------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| gene_synonym = Aliases                                                                                                | Symbol             | CDK2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|                                                                                                                       | Description        | cyclin dependent kinase 2                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| chromosome_element_id = "chr12"<br>start_position = "55966769"<br>end_position = "55972784"<br>hg_identifier = "hg38" | Accessions         | 1017 (NCBI Gene)<br>ENSG00000123374 (Ensembl)<br>P24941 (UniProt)<br>118953 (OMIM)<br>74409 (HomoloGene)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|                                                                                                                       | Aliases            | CDKN2, p33 <sup>CDK2</sup>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|                                                                                                                       | Genome Location    | chr12:55966769-55972784 (hg38)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|                                                                                                                       | Molecular Function | magnesium ion binding (GO:0000287)<br>protein serine/threonine kinase activity (GO:0004674)<br>cyclin-dependent protein serine/threonine kinase activity (GO:0004683)<br>cyclin-dependent protein serine/threonine kinase activity (GO:0004683)<br>cyclin-dependent protein serine/threonine kinase activity (GO:0004683)<br>protein binding (GO:0005515)<br>ATP binding (GO:0005524)<br>protein domain specific binding (GO:0019904)<br>cyclin binding (GO:0030332)<br>cyclin binding (GO:0030332)<br>cyclin binding (GO:0030332)<br>histone kinase activity (GO:0035173)<br>cyclin-dependent protein kinase activity (GO:0097472) |
|                                                                                                                       | Biological Process | G1/S transition of mitotic cell cycle (GO:0000082)<br>G1/S transition of mitotic cell cycle (GO:0000082)<br>G2/M transition of mitotic cell cycle (GO:0000086)<br>G2/M transition of mitotic cell cycle (GO:0000086)<br>positive regulation of transcription by RNA polymerase II (GO:0000086)                                                                                                                                                                                                                                                                                                                                      |

Como se mencionó anteriormente la nomenclatura utilizada en las BDGH no siempre es la misma que maneja el investigador por lo que en diferentes casos se puede omitir información relevante a su búsqueda por la omisión o desconocimiento de cierta nomenclatura. Para explicar usamos el ejemplo con la BDG "SitEX" (16), en la cual los nombres presentados en la BD no son los mismos a los indicados en el ECGH, por lo que es importante el conocimiento del tema para determinar cada uno de los elementos que contiene la base de datos. En la

Ilustración 5, se aprecia por colores la información válida, que en este caso es Strand (hélice del ADN) en color verde, id del gen color rojo, id de transcripción color amarillo y nombre de la proteína color rosado.

### Ilustración 5. Forma de presentación de la información para los elementos de cromosoma y transcripciones

| Chains                                            |                                                                                                   |
|---------------------------------------------------|---------------------------------------------------------------------------------------------------|
| ChainName:                                        | A                                                                                                 |
| EnsGene:                                          | ENSG00000087085                                                                                   |
| EnsTranscript:                                    | ENST00000241069                                                                                   |
| EnsProtein:                                       | ENSP00000241069                                                                                   |
| ENSMolecule:                                      | acetylcholinesterase (Yt blood group) [Source:HGNC Symbol;Acc:HGNC:108] (List of exons and sites) |
| Site positions, AA:                               | 378_381                                                                                           |
| Discontinuity in sequence:                        | 0.500000                                                                                          |
| Discontinuity in exon structure:                  | 0.000000                                                                                          |
| Average sequence identity in site:                | 0.951220                                                                                          |
| Average Kabat conservation score:                 | 2.050000                                                                                          |
| Exon structure variation in functional site area: | 0.000000                                                                                          |

Tomando la base de datos genómica de "ACVR1" (17) se puede apreciar en la Ilustración 6 que lo enmarcado en color rojo pertenece en el ECGH a Strand (hélice del ADN), pero en este caso se encuentra indicado como "negative strand". De esta manera la extracción automática de información puede ser obsoleta si no se considera lo descrito entre paréntesis.

### Ilustración 6. Detalles de información que deben ser consideradas para la extracción automática

**Gene**  
ACVR1

- Gene view
- Overview
- External links
- Drug resistance
- Tissue distribution
- Genome browser
- Mutation distribution
- Variants
- References

Reset case

**Search**

Search COSMIC...

**Filters**

Show advanced filters

Range Show input fields

1 510

**Overview**

This section gives an overview of ACVR1, along with links to any related data and resources.

Census gene Curated gene Mouse gene Hallmark gene

**COSMIC gene** ACVR1 (COSG37)

**Genomic coordinates** 2:157737531..157799493 (negative strand)

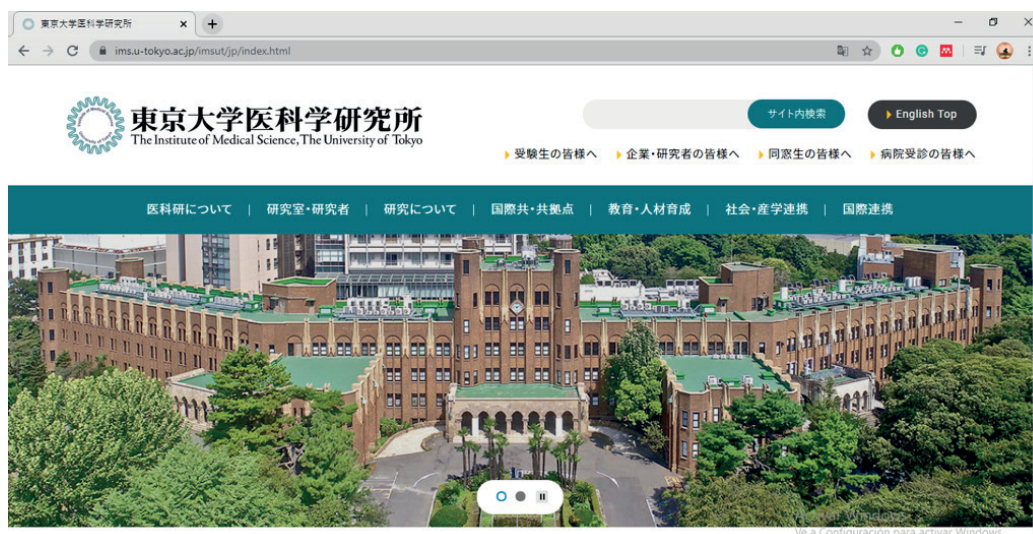
**Synonyms** ACVRLK2, CCDS2206.1, Q04771, ENSG00000115170

**COSMIC-3D** There are 10 structures for ACVR1. View them in COSMIC-3D.

**Number of samples** 59519 unique samples  
255 unique samples with mutations

El idioma que utilizan las diferentes BDG es uno de los problemas que suelen evidenciar los investigadores, esto conlleva al descarte o extracción errónea de la información por el desconocimiento del idioma. Para evidenciar este problema se toma la base de datos genómica "Comparasite" (18), en la que se puede apreciar (Ver Ilustración 7) que se encuentra en otro idioma, pero el contenido también se lo puede visualizar en inglés.

## Ilustración 7. Sitio web presentado en 2 lenguajes



Así mismo el contenido presentado en la base de datos "CleanEST" (19) se encuentra en un idioma diferente al inglés en el que usualmente se encuentra todo el material genético. En algunos sitios web se dispone de la traducción propia dentro del sitio, pero en este caso no se obtiene traducción alguna (Ver Ilustración 8), por lo que se genera una confusión en el contenido que el sitio web presenta.

## Ilustración 8. Sitio web presentado en lenguaje nativo

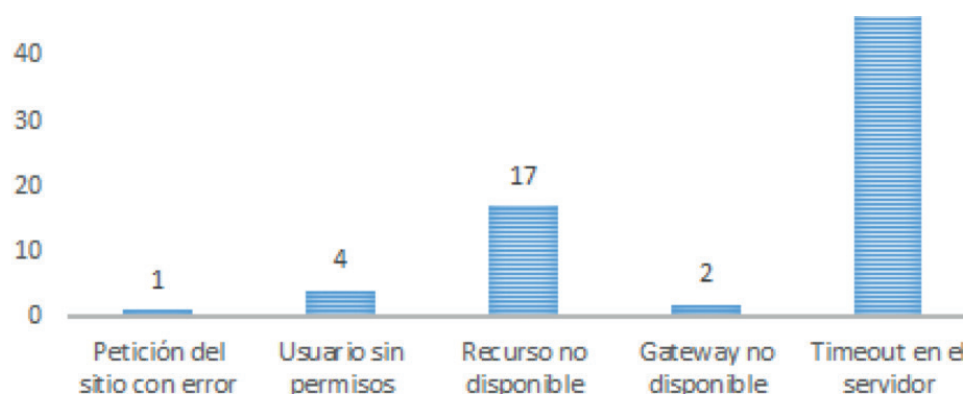


En cuanto al manejo de diferentes idiomas para las BDG se puede generar scripts de detección y idioma y posterior traducción de la información que se está manejando para su correcta extracción y manejo de esta.

## Mantenibilidad

Los sistemas de información genómica no dan un mantenimiento constante como se puede ver en el trabajo realizado por Solís (9), en donde se exploró 761 bases de datos genómicas públicas que tratan información sobre humanos solamente el 90.14% (686 DBGH) se puede tener acceso, mientras que el 9.86% (75 DBGH) no se puede acceder a los recursos. Entre las posibles causas de no acceso (ver imagen 9) se tiene 55 casos un el servidor donde están alojados los recursos no da una respuesta (time out server).

**Ilustración 9. Causas de no conectividad**



Así mismo revisando la información que se indica en cada una de las BDGH se tiene que el 85.41% (650 BDG) no han sido actualizadas versus un 14.59% (111 BDG) que, si lo han hecho, esto puede causar a nivel de investigadores desconfianza de la información publicada.

Cabe indicar que hay dos puntos muy relevantes que indirectamente colaboran con el “Caos genómico”, uno de ellos es que los diferentes sitios web no han recibido mantenimiento, siendo otro punto que en algunos casos los servidores han dejado de funcionar y algunas páginas web han perdido incluso su dominio por lo que no se obtiene ni siquiera una portada o datos informativos sobre el proyecto que trabajaban. Este es un punto clave ya que algunos sitios podrían dar información valiosa, pero al perderse el rastro simplemente se omiten e ignora la base de datos genómica.

Otro punto relevante es el hecho de que algunos sitios web en donde las bases de datos genómicas eran públicas en un inicio, ahora se han convertido en sitios de pago o a su vez se requiere de una invitación para la creación de una cuenta de acceso.

Ante el acceso libre a las BDG que pueden cambiar durante el tiempo, no se puede garantizar una solución permanente puesto que en algunas circunstancias es debido a que sus proyectos terminaron y no se dispone de un sitio para alojar la información y estas desaparecen, o quizás dejaron de ser de acceso público y ahora son de pago. Sin embargo, la creación de scripts que indiquen el estado de estas DBG puede ser un indicador en el momento de la extracción de información.

## CONCLUSIONES

La información publicada en los diferentes repositorios o base de datos que se encuentran en la web por varios grupos de investigación o consorcios formados a lo largo del tiempo han permitido el acceso a dicho material por parte de investigadores del área médica y personal entendido en el tema. Sin embargo, la magnitud de información ha generado malestar ya sea porque la información genómica no se ha podido manejar de manera óptima de manera que se pueda generar reportes con datos significativos; es aquí donde con ayuda de la informática se puede resolver algunos de estos inconvenientes y mitigar el caos genómico.

Ante los problemas tratados, en este trabajo se emiten criterios como la conformación de una base genérica de información genómica a nivel de humanos. Para ello se ha propuesto la utilización de un esquema conceptual del genoma humano que permite la extracción de datos mínima, pero a su vez necesaria para comprender la estructura y manejo de información genómica, además que esta base de conocimiento genera un punto de partida para la estandarización de la información genómica humana. La utilización de un esquema conceptual permite tener una versatilidad en el manejo de la información en este caso enfocado al genoma humano.

Si bien los atributos requeridos por investigadores o médicos son adaptados a medida que se adquieren nuevos conocimientos en base a la investigación científica. Si se requiere incorporar nueva información, bajo el modelo conceptual del genoma humano se permite realizarlo sin inconvenientes, al igual que si se requiere omitir o eliminar. Como medida para evitar el caos genómico muchos de los investigadores se han familiarizado en el manejo de al menos 3 o 4 sistemas de información para lograr rapidez en la comprensión de los datos. Dependiendo del área en la que se quiere aplicar los datos disponibles, se debería seleccionar un sistema de información genómico amplio y diverso, que se enlace con otras bases de datos delimitados o específicos y que sea compatible con la disponibilidad de tecnología para experimentar. Como ejemplo de un sistema de información que considera algunas especies pues se selecciona *H. sapiens* o Human, se busca la información genómica de interés, posterior enlazar con una base de datos específica por ejemplo cáncer de mama; y si va a experimentar analizar la tecnología a disposición (microarray, secuenciación, PCR, etc), disponibilidad de reactivos e insumos en el mercado.

Los sistemas de información genómica deben consensuar para integrar los datos genómicos humanos, manejar una nomenclatura común para volverlos más eficientes al momento de gestionar la información y disponer de la opción de idiomas o traductores

El caos genómico no se puede solucionar, pero tampoco se lo puede ignorar, la propuesta es mitigarlo con la ayuda de herramientas informáticas para manejo de estructuras y grandes cantidades de datos con la finalidad de que los datos se conviertan en información útil para todas las personas vinculadas con el mundo genómico especialmente con los investigadores, personal de laboratorios genéticos y médicos en general. Sin embargo, mientras no se lo pueda estandarizar y mantener un orden y secuencia de la información publicada en los diferentes sitios web mediante los diferentes recursos informáticos no se podrá mantener un control de todo lo que se dispone.

## REFERENCIAS BIBLIOGRÁFICAS

- Watson JD. The human genome project: Past, present, and future. *Science*. 1990;248(4951):44–9.
2. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. Vol. 422, *Nature*. Nature Publishing Group; 2003. p. 835–47.
3. 10 años de la secuenciación del genoma humano: Encuentro entre el imaginario y la realidad [Internet]. [cited 2020 Apr 27]. Available from: [http://www.scielo.org.co/scielo.php?pid=S0123-34752013000100001&script=sci\\_arttext&tlng=en](http://www.scielo.org.co/scielo.php?pid=S0123-34752013000100001&script=sci_arttext&tlng=en)
4. Craig Venter J, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001 Feb 16;291(5507):1304–51.
5. von Knebel Doeberitz M. New markers for cervical dysplasia to visualise the genomic chaos created by aberrant oncogenic papillomavirus infections. *European Journal of Cancer*. 2002 Nov 1;38(17):2229–42.
6. Lorenz S, Barøy T, Sun J, Nome T, Vodák D, Bryne JC, et al. Unscrambling the genomic chaos of osteosarcoma reveals extensive transcript fusion, recurrent rearrangements and frequent novel TP53 aberrations. *Oncotarget*. 2016;7(5):5273–88.
7. Pardo A. El genoma humano. Límites y perspectivas en el avance de la medicina [Internet]. Vol. 40, *Arch Bronconeumol*. México; 2004 [cited 2020 Apr 27]. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7131294/pdf/main.pdf>
8. Reyes Román JF. DISEÑO Y DESARROLLO DE UN SISTEMA DE INFORMACIÓN GENÓMICA BASADO EN UN MODELO CONCEPTUAL HOLÍSTICO DEL GENOMA HUMANO. 2018 Mar 22;
9. Alexandra V, Cabrera S. Exploración de Bases de Datos Genómicas Dirigida por Modelos Conceptuales Septiembre 2018. 2019 Jan.
10. Vidal-Casero M del C. EL PROYECTO GENOMA HUMANO. SUS VENTAJAS, SUS INCONVENIENTES Y SUS PROBLEMAS ÉTICOS [Internet]. Valencia; 2001 [cited 2020 Apr 28]. Available from: <http://aebioetica.org/revistas/2001/3/46/393.pdf>
11. UCSC Genome Browser Home [Internet]. [cited 2020 Apr 28]. Available from: <https://genome.ucsc.edu/>
12. PKD Mutation Database [Internet]. [cited 2020 Apr 28]. Available from: [https://pkdb.mayo.edu/cgi-bin/v2\\_display\\_mutations.cgi?apkd\\_mode=PROD](https://pkdb.mayo.edu/cgi-bin/v2_display_mutations.cgi?apkd_mode=PROD)
13. El S, De R, Migraciones L, De AT, Genoma N. LA GENÉTICA DE LAS MIGRACIONES HUMANAS. *MÈTODE Science Studies Journal*. 2014;4.
14. TransmiR v2.0 [Internet]. [cited 2020 Apr 28]. Available from: <http://www.cuilab.cn/transmir>
15. BioGPS - your Gene Portal System [Internet]. [cited 2020 Apr 28]. Available from: <http://biogps.org/#goto=welcome>
16. SitEX [Internet]. [cited 2020 Apr 28]. Available from: <http://www-bionet.sccc.ru/sitex/>

index.php?siteid=238202

17. ACVR1 Gene - Somatic Mutations in Cancer [Internet]. [cited 2020 Apr 28]. Available from: <https://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=ACVR1>
18. 東京大学医科学研究所 [Internet]. [cited 2020 Apr 28]. Available from: <https://www.ims.u-tokyo.ac.jp/imsut/jp/>
19. 국가생명연구자원정보센터 KOBIC [Internet]. [cited 2020 Apr 28]. Available from: <https://www.kobic.re.kr/>

#### **CONTRIBUCIÓN DE LOS AUTORES.**

Vanessa Solís-Cabrera (VSC), Lourdes Viñansaca (LV), Juan José Sáenz Peñafiel (JJSP): recolección de los datos, revisión bibliográfica, escritura del manuscrito y análisis crítico del

manuscrito.

#### **INFORMACIÓN DE LOS AUTORES.**

**Vanessa Solís-Cabrera:** Estudiante doctoral Universitat Politècnica de València Ingeniera de Sistemas. Investigadora Universidad de Cuenca-Ecuador. Magister en Gestión Estratégica de Tecnologías de Información. Máster Universitario en Ingeniería y Tecnología de Sistemas Software.

**Lourdes Viñansaca:** Docente Universidad de Cuenca-Ecuador. Licenciada en laboratorio clínico. Máster en docencia universitaria.

**Juan-José Sáenz-Peñafiel:** Estudiante doctoral Universitat Politècnica de València. Ingeniero de Sistemas. Máster Universitario en Ingeniería de Computadores y Redes. **FINANCIAMIENTO.**

La investigación fue financiada por los autores

#### **AGRADECIMIENTO.**

El presente estudio fue realizado dentro del grupo PROS

#### **DECLARACIÓN DE INTERESES.**

Los autores declaran no tener conflicto de intereses.

#### **DISPONIBILIDAD DE DATOS:**

Todos los datos se encuentran a disposición de los lectores.

#### **AUTORIZACIÓN PARA LA PUBLICACIÓN**

Todos los autores autorizan su publicación.