

CRISPR

Science is such a broad based topic which encompasses information valuable across industry's. It was so when at Danisco, a Danish food production company, one of their researchers discovered CRISPR. When searching the cellular structure of Streptococcus Thermophilus, a bacterium used to make yogurts and cheeses, and the ability for certain viruses to infect and damage the quality of our foods the observance of CRISPR sequencing arose. It is found to grant Streptococcus Thermophilus bacteria with an adaptive immunity against viral invasions. It was then transformed into a plausible reality and use for animal and human biological systems.

At that time, CRISPR became a new technology which facilitates making specific changes in the DNA of humans, other animals, and plants. CRISPR is an acronym which stands for 'Clustered Regularly Interspaced Short Palindromic Repeat'. It is in reference to the 'repeated' DNA sequence found in genomes of bacteria and other microorganism. These sequences are an important aspect of the immune systems as it protects organisms health and well being. When these sequences are invaded by a 'virus', it is CRISPR that come between cells nature consistency and the invading virus. It seeks to destroy the genome of the invading virus.

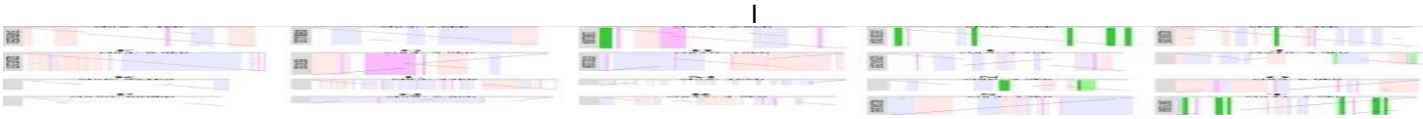
One cellular element identified as 'Cas 9 protein' is key to CRISPR function. This protein can be programmed to 'find' and bind to almost any desired target sequence RNA to guide it in it's search of the 20 – DNA letter long sequence that matches part of the guide RNA. Research suggest that there are approximately six billion letters packed into each of our cells. This Cas 9 cuts DNA at it's target point. When it's 'cut' is repaired, mutations are introduced that typically 'disable' a gene.

A genome is the haploid set of chromosomes with the genes they contain. The genome is the area of focus for destroying invading viruses. As CRISPR imitates this DNA sequence of the 'invading' virus genome, it's ability to replicate and spread is disabled and keeps the virus from growing. This is a 'repair' state of the cell which relies upon CRISPR sequencing being transcribed and processed to generate 'short' CRISPR RNA molecules. These molecules lead to bacterial molecular machinery as a matching target sequence in the invading virus which disengages the viral DNA sequence and destroys the virus.

Some regions of the cell house 'short' DNA repeats known as black diamond and spacers [colored boxes]. This allows for unseen viruses to infect bacterium as non-spacers are incorporated amongst existing spacers. Spacers have a 'genetic memory' of previous infections which message CRISPR to eliminate the DNA sequence and destroy it again.

This occurs with unseen virus attacks previous tracked by researchers whom consider that as new spacers are made and added to the chain of spaces and repeats three phases take place:

1. Adaptation : 3 basic steps of protection DNA & virus inserted in CRISPR sequence is new spacers
2. Production : CRISPR repeats & spacers in bacteria DNA is copied [transcribed] to RNA then cut into 'short' pieces as CRISPR RNA
3. Targeting : DNA II [double helix] and RNA I [single strand]



This same technologic body of knowledge in CRISPR research, furthered it's use having branched into it's ability to study 'gene editing'. Gene Editing is the ability for researchers and scientist to 'repair' templates with specific change in sequencing added to the cell and incorporated in the DNA during a 'repair' process of targeted DNA which is altered to carry a new sequence changing or correcting it's functionality.

'Synthesizing' short RNA molecules that match a specific DNA sequence, as example 'human cells' guide RNA which shuttle molecular machinery to the intended DNA target, once localized, this machinery can 'silence' a gene or even change the sequence of a gene hence gene editing. During this 'silencing' cells attempt to repair the broken DNA, but does so with 'errors' that disrupt the gene 'silencing' it. If disease invades our biological cells, the DNA's short RNA component message the CRISPR onto the intended DNA target but change it's sequence with error then silencing it and it's ability to spread or function. In medicine, CRISPR can be used to correct a 'mutant' gene and reverse disease symptoms in a living animal through correction of the mutant gene with it's proper 'sequence'.

It too enables more specific antibiotic manufacturing which target only the disease causing bacterial strains while sparing beneficial bacteria.

Future use is comprised of:

1. Changing genomes of our children
2. Fingerprinting cells
3. Logging
4. Directing evolution & creating gene drives

DNA copying

Researchers have sought out the comparative study of 'DNA repair genes' utilizing human population screenings as they seek to impact protein function. Through this means of study, over 520 different amino acid substitutions variants have been identified using 91 human DNA repair gene for sequence variation.

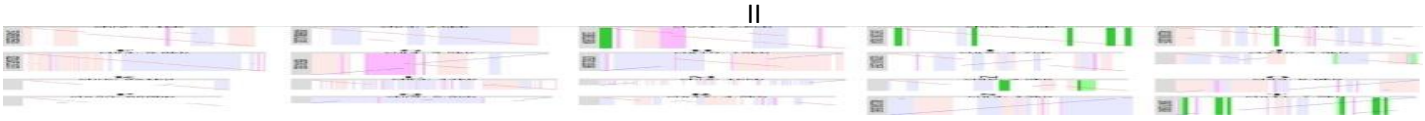
There were over 1.2 million single nucleotide polymorphisms [SNPs] identified as were the many common 'missense' variants in genes with potential fractioning those susceptible to common human disease. SNP's are variations at a single position in a DNA sequence among individuals which may lead to variations in amino acid sequences. [protein] Genes encode proteins with roles in the repair of damaged DNA sequence variations having over five different amino acid substitutes per gene within the sequencing of 84 DNA repair genes of 90 healthy individuals.

There are two forms of study of genetic protein impact:

1. Polymorphic

2. SIFT [Sorting Intolerant From Tolerant]

The first, Polymorphic variants impact protein and the repair pathway activity. These are comprised of the association of several polymorphic variants with 'reduced' DNA repair capacity which result in the elevation of biological disease. For instance, humans with such 'reduced' DNA repair ability will have a increase in 'cancer' cell replication. Methodology of variant identity incorporated developing approaches and protocols for identifying, control the number and samples of



genes screened [human] and analyzing the predicted impact of activity by SIFT and PolyPhen in comparative use of selected databases for preregistered sequencing.

Algorithms for both sequence forms are from conservation over evolutionary time, the physical and chemical properties of the exchanged residues and or protein structural domain information.

1. SIFT algorithm sequence homology among related genes and domains over evolutionary time and residue characteristics in substitution amino acids prediction.
   SIFT is a sequence homology based tool that sorts intolerant from tolerant amino acid substitutions.

2. PolyPhen [Polymorphism Phenotyping] algorithm also incorporates sequence and the nature of amino residues involved.
   PolyPhen score is the probability that a substance is damaging.

This algorithm process values the location of the substitution within identified functional domains and known structure with protein features as viewed in the SwissPro database. There are several databases for referencing identified sequencing proteins and other aspects of cellular functionality. Swisspro, GenBank, BAC, PAC, and YAC [Yeast artificial chromosomes] libraries are resources of sequencing data.

Once the specific algorithms for SIFT and PolyPhen were formed, they were used to a set of 523 amino acid substitutions in screening of 91 DNA repair genes from human sampling. There scores were categorized into three groups: (1) Borderline, (2) Damaging, and (3) Tolerated.

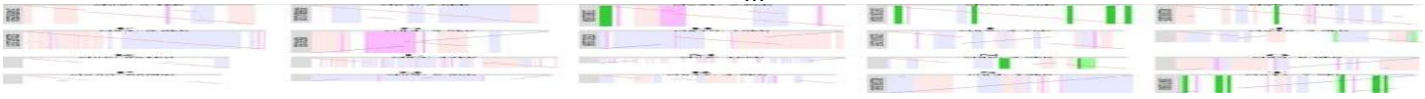
Both algorithms identify substitutions impacting enzymatic activity of DNA repair proteins, score obtained from 26 amino acid substitution variants of APEX1. This is 'mutagenesis' the study of enzymes mechanisms. Here researchers found a 'benchmarking' ability for the SIFT and PolyPhen algorithms in confirmation that both are capable of discriminating between variants with minimal residual activity and variants with wild-type activity. 75% of the amino acid substitutions in a subset of the genes occurred at residues that were identical in humans and mice.

Gene damage within our cells has outcomes such as Bloom's disease, colon cancer, breast and ovarian cancer syndrome, genetic disorder, or Fanconi Anemia as there is gene mutation. The repair of damaged DNA require proteins to function in multi-protein complexes. The reality that there are a substantial number of potentially relevant variations in a repair pathway is a 'hindrance' to this aspect of research.

*Employed a broader definition of agreement with the cells adjacent to the 'diagonal' cells

The Human Genome Project had its first landmark in sequencing of the euchromatic portion of a human chromosome in 1999. All chromo-human have two arms a long and a short [L & S]. This process proved that there exist no evidence of protein coding genes on the 'short' arm [cellular] however, the 'long' arm of the chromosomes are rich in genes compared with other chromosomes.

III



Short Arm = arm symbolized as a "p" from the word 'petit' meaning small.

Long Arm = arm is symbolized as a "q" for long arm as it 'follows' "p" in the alphabet. This part of the chromosomes are attached to the 'spindle' during cellular division.

Identifying genomic clones as 'clone maps' of chromosomes were constructed using a constructive process:

1. Clone by clone sequencing the genome is broken up into large long chunks [base pairs]
2. The location of these chunks on the chromosome is mapped.
3. Chunks are placed in Bacterial Artificial Chromosomes [BAC] and put inside bacterial cells to grow.
4. DNA is copied within the 'division' phase within the cell.
5. DNA in clones aare broken down into small overlapping fragments
6. Fragments are put into vector known as DNA sequence
7. Sequence range from known to unknown
8. Area of overlap is Identified to reform those large chunks
9. The 'Map' is used to reconstruct large chunks back into chromosomes

The 'regions' of sequencing are separated by a central contig of 23MB. The ability to capture prediction accuracy varies between different regions of sequence and need more reliable 'rules' to be established based on ab initio methods. Few of the gene structures are completely correct having 95 percentile of the prediction are partially correct. 94% of the annotated genes are at least partially detected using the Genscan gene prediction. 20% have all exons predicted exactly.

60% of human genes have distinctive CpG island sequences at their 5' ends which can be used for identifying potential genes. This means uses a combination of similarity searches against all available DNA and protein databases as well as a series of ab initio predictions. Those 'masked' sequences were compared to public domain DNA and protein databases by similarity searches using the 'blast' family of programs.

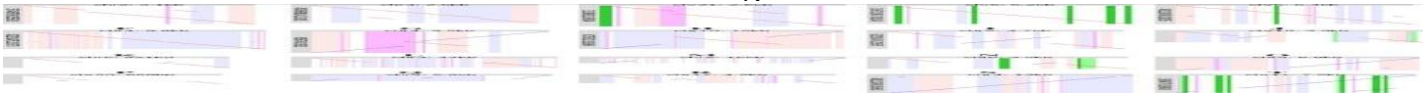
The outcomes are grouped in accordance to the evidence that was used to identify them:

a. Known Genes are those genes identical to known gene or protein sequence
b. Related Genes are those homologous genes or containing a region of similarity
c. Pseudo-Genes are homologous to a known gene or protein having a disrupted open reading frame

Gene Prediction Program

Gene prediction depends upon the use of 'putative genes' and it identifies 'regions' of genomic DNA that 'encode' genes.

There are two classes of methods which are adapted for (a) similarity searches or (b) *ab initio* prediction for finding the location of 'protein' coding regions. Gene prediction is focused within the coding regions, shadow coding regions, and non-coding DNA of cells. The 'Exon' protein coding tissue is the section of a DNA or RNA molecule containing information coding for a protein or peptide sequence.



During this process, identified sequence (s) are compared with library databases of sequences. There exist 'generational' technological advancement in evaluating the evolution of such technology and bio-database compilation.

1] First Generation: could not predict 'exon' locations [Ecoparse, GenMark, Test code, GRAIL]

2] Second Generation: combined splice signal & coding region ID to predict Exons. These sequences did not assemble exons into complete genes. [SORFIND, XPound]

3] Third Generation: predicting complete gene, however, poor performance programs. [Gene ID, Gene Parser8, GenLong, FGENEH]

4] Fourth Generation: improved accuracy. [GENSCAN, AUGUSTUS]

Gene Prediction Methodology's

A. Ab initio

This ab initio method relies on (1) signal sensors and (2) content sensors. Content sensor are used by exons and allow for coding sequence distinction from non-coding sequences.

B. Blast family of programs

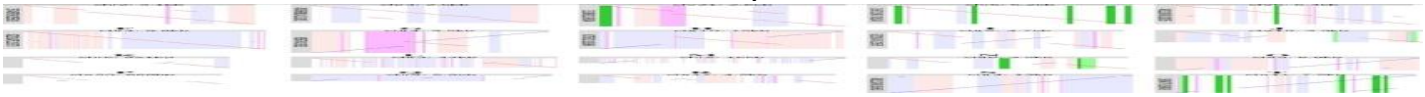
The Blast (basic local alignment search tool) is an algorithm which compares primary bio-sequencing information of amino acid [proteins] or nucleotides DNA and RNA.

In this new era of genetic superfection, we are learning of the scientific ability to prescreen embryos for any number of traits. These first assumptions in value include 'in vitro fertilization' for primarily parental and insurance uses and pre-school and college scoring DNA for admission purposes. Dating sites may even begin asserting the us of 'genetic' profiling for matching depth.

The "Predictive Genetic Analysis" [PGA] provide Polygenetic Scores [PGS] which comprises of a single number which is indicative of measurable genetic influence on given traits such as embryo height, depression, diabetes, cognitive ability, or even skin tone.

No matter how relevant 'genetics' are to controlling and improving the health and intelligencia of a community, there is a balance of inequities created against potential lives saved or improved. In 2008, the Genetic Information Non-discrimination Act was adapted which states: health insurers and employers are barred from using genetic information in their decisions however, this law does not encompass schools, dating, sperm-banks, fertility uses, etc….. Nor have we deeply understood the ability to pre-determine diseases such a schizophrenia, bi-polar, heart disease, downs syndrome, as a means to approved abortive practices.

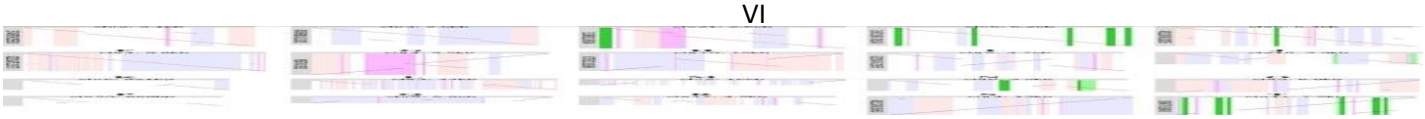
There is also consideration of the may obscure health, mental, or other inequities by race as there are fewer studies with large pools of Black, Hispanic, and other culture participants.



The conclusion of this advanced research is on a never-ending path of refining targeted methods and evolving the many plausible uses of CRISPR – Cas 9 Protein in human health and predictive and corrective gene editing. As science expands into this era, we've learned of the latest updates.

1. Targeted mutations to the genome can now be recognized by splitting specific enzymes and then triggering them to reconstitute.
2. There is the use of 'base editors', the latest and most effective way to achieve precise gene editing, to locate specific DNA targets and DNA deaminase enzymes to modify and mutate the target.
3. Base editors C:G base pairs in DNA can be mutated to T:A or A:T base pairs can be turned to G:C.
4. DNA deaminases can be divided into two pieces which can be put back together using a small cell permeable molecule called "rapamycin".
5. This "rapamycin" can be introduced and lay dormant in a cell until the small molecule is added & then the 'base editing' complex is turned on to alter the genome.
6. There is no way to trigger mutations at specific times or keep the editor in check to prevent undesired mutations.

Farther review of information relative to CRISPR can be found in video review and chats online within the Youtube and TED'S Talks platforms.



REFERENCES

Gardner, Alex. New Gene-Editing Technique Offers Scientist Ability to "Turn On" Enzymes That Cause DNA Base Mutations. Penn Medicine News, October 28, 2021. www.pennmedicine.org

Le Page, Michael. New Scientist: CRISPR a technology that can be used to edit genes. CRISPR to alter plants, animals in our farms, gardens or homes may have been in near future, Newscientist.com November 13, 2021

I. Dunham, N. Shimizu, B. A. Roe, S. Chissoe et al. The DNA sequence of human chromosomes 22. NATURE/VOL 402/2, December 1999

Science News In The New Harvard University School of Arts and Science, 'CRISPR: A game-changing genetic engineering technique, July 31, 2014

Thegenehome.com, November 13, 2021

T. Xi I. M. Jones and H. W. Mohrenweiser. Variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Issued in 2004. Genomics [2004] 83, 970 – 079. Biology and Biotechnology Research Program. Lawrence Livermore National Laboratory, Livermore CA.

www.yourgenome.org. What is clone by clone Sequencing, 2021-07-21. November 29, 2021

Zhou Wand, Yazhu Chen, and Yixue Li. Genomics, Proteomics Bioinformatics: A Brief Review of Computational Genes Prediction Methods. Ncbi.nlm.nih-gov/ 2004 Nov

