Edgar **Fragoso**

Data Science · Bioinformatics

(+52) 55-211-849-86 | edgarfragosogarcia@gmail.com.mx | edgar-omar-fragoso-garcia-69006092

Summary

Bioinformatics and Data Science are multidisciplinary topics, is necessary to have abilities in programming, statistics, mathematics, and biology. 7+ years experience in these areas where I have worked for different companies where the common objective is to generate knowledge through the data for desitions making. The data are in continual growth, there are new manners of resolve problems and visualize them, which drives me to be an autodidact. Highly analytical and skilled bioinformatics analyst always ready to investigate and manipulate complex material. I get involved in multidisciplinary teams.

Education

IPN-UPIBI(UNIDAD PROFESIONAL INTERDISCIPLINARIA DE BIOTECNOLOGIA)

CDMX, México

B.S. in Biotechnology Engineering — *Aug. 2009 - Aug. 2013*

- Characterization of pathogenic organisms using NGS (454 GS-FLX and Ion PGM) and bioinformatics tools (DB: ENTREZ, Pubmed, Warwick, Silva, RDP, KRAKEN, OneCodex, Software: CLC-BIO, BIONUMERICS, BLAST, FASTQC, Trimmomatic, Base Space, SAMTOOLS, PICARD, IDBA, BOWTIE, Mauve, Mugsy, OPERATIVE SYSTEM: Fedora, Ubuntu, Debian, PROGRAMING LANGUAGE: shell and python).

Work Experience

Somos Ancestria.

SF, U.S.A

Senior Bioinformatics Developer & Data Scientist — *Oct. 2020 - Present*

- Built fully automated a pipeline with a shell script and python for the calculate individual ancestry compositions using tools as Plink and Admixture the pipe is mount inside a virtual machine with Centos on AWS-EC2, the pipeline returns a json file for insert the values at MongoDB.
- Panel design of reference, the open data was taken from 1000 Genomes Project and HGDP to be merged with a customized panel of Latin people, and then apply different quality control, the final matrix with SNPs and subjects was defined through cross-validation.
- Developed a parser in python for the standardization of files with genetic data that subsequent will be used by snappy and haplogrep for calculating a haplogroup.
- Built database with information of haplogroups as age and location the data was drawn out with techniques of web scraping in python, the data are deployed in an interface for the user to interact with the map and a tree that show the relation father-child.
- Categorization of data coming from surveys using libraries like Pandas, Numpy and Plotly, for downstream analysis like GWAS and population structure.

GENOLIFE

Sonora, México

Bioinformatics Assessor — *Nov. 2020 - Feb. 2021*

- Advice development of a pipeline for calculating Polygenic Risk Score in Mexican. The beta scores for genetic risk variants were taken from SIGMA project and UK-biobank to calculate a normal distribution with Plink.

Código46

Cuernavaca, México

Senior Bioinformatics Developer & System Administrator — *Jul. 2017 - Nov. 2020*

- Built fully automated a pipeline with a shell script for calculating individual ancestry compositions using tools as Plink and Admixture. the pipe returns a flat file to be insert at SQL database.
- Built Panel of reference, the different subjects that represent one specific population were drawn out of 1000 Genomes Project, HGDP and Estonian Biocentre and merged for apply filters as a threshold of call rate, maf, equilibrium hardy-weinberg, identity by descent, linkage disequilibrium and remove outliers.
- Developed a matrix with allele frequencies and the ancestral portion was calculated by Admixture software using cross-validation and PCA to determine the best clusters.
- Built fully automated a pipeline with a shell script for calculating the relationship between individuals using the IBD value with several SNPs and subjects, the above to infer a possible relation of paternity.
- Participation of XLIV NATIONAL CONGRESS OF HUMAN GENETICS carried out at Chiapas where I expose my job named " CONTRIBUTIONS OF THE FIRST AMERICANS IN NATIVE POPULATIONS AND CONTEMPORARY IN MEXICO".
- Assist in activities of investigation to the project named "Screening for metabolic syndrome and obesity from genetic data of the adult population of the state of Morelos", my main functions were to review the instruments that would be filled by patients, once processed the samples by Illumina iScan System, I builted a EGT new file with > 1000 samples to improve GenCall Score Cutoff and Call Rate.
- Assist in activities like mounting the network infrastructure, configure IP addresses for the different equipment of laboratory-like Tecan, Server of Illumina (LIMS), and Iscan, open ports with differents protocol ssh and FTP for share files, instance deployment in AWS-EC2 with AMI of Centos for processing data and mount different pipelines. Installation of RedHat Enterprise in Power Dell in RAID5. Configuration of KVM for mount virtual machines where live the frameworks are developed by BC-Platforms. Configuration of Site Recovery service of Azure for migrating the ISO from AWS-EC2.

SENASICA-SAGARPA *Tecámac, México*

Bioinformatics Assessor *Jun. 2018 - Jul. 2018*

- Built fully automated a pipeline with software CLC GENOMICS WORKBENCH for the analysis of sequences of ADN come from platform Miseq of a virus of Influenza A, applied different QC as the length of reads, GC content and PHRED distribution, the type of the read paired-end were assembled using a different algorithm like IDBA and CLC de novo assembly for comparing the distribution of contig (relation of N50 vs insert size), the contigs were compared with sequences in NCBI using blast algorithm for determinate the type of hemagglutinin and neuraminidase. The phylogenetic trees showed that segments were related to strains of influenza A already found in Mexico.

Winter Genomics *CDMX, México*

Bioinfomratics Developer *Jun. 2017 - Jul. 2017*

- Analysis of coverage using the best practices GATK and Platinum Genomes for generating *.bed files with possible SNPs or INDEL in Ameridians populations.

SENASICA-SAGARPA. *Tecámac, México*

Bioinformatics Developer & System Administrator *Sep. 2013 - Feb. 2017*

- Developed a pipeline for Genomic Characterization of Salmonella Enterica, Escherichia Coli, Listeria Monocytogenes using data of sequencing came from platforms like Illumina (Nextseq), Ion Torrent (PGM), 454 GS-FLX.
- Developed database with sequences of housekeeping genes to implemented the technique of Multilocus sequence typing.
- Developed a pipeline for genomic characterization of Pathogenic organisms that can not be identified with traditional techniques of bacteriology using 16S ribosomal.
- Implementing of software SeqSero for the determinate the serotype and Antigenic formula of strain Salmonella Enterica, Escherichia Coli and Klebsiella the result was showed to WHO Collaborating Centre External Quality Assurance System (EQAS) 2016.
- Participation in the trial organized by Global Microbial Identifier who gather and shared the results obtained of different laboratories around the world both wet-lab (metrics of ADN extraction and status of the sequencing run) and dry-lab (metrics of coverage, the status of assemblies, length of reads). The pathogenic organisms were identified using 16S-RNA (specie), MLST(serotype), and ARG-ANNOT (Antibiotic Resistance).
- Developed of molecular markers using Phylomark that use complete genomes to identify conserved phylogenetic markers for typify of Salmonella and E. coli the markers were used by endpoint PCR for epidemiological surveillance.
- Support to organizational activities to carry out the 1° and 2° International Symposium on Next-Generation Sequencing in the installations of SENASICA-SAGARPA "Unidad Integral de Servicio, Diagnóstico y Constatación, Tecámac, Edo. México. My presentation was about "Characterization of pathogenic organisms using next-generation sequencing and bioinformatics".
- Presentation of poster in Latin food 2016 with the name "Genomics and Phylogeographic of Samonella in Mexico" carry out in Cancún Qro, XXX Congreso Nacional de Bioquímica carry out in Guadalajara Jal. "Characterization of pathogenic organisms using next generation sequencing and bioinformatics".
- Follow up the meeting Pulsenet América latina y el Caribe 2015, where was presented the project "Genómica y Filogeografía de Salmonella en México" in cooperation with COFEPRIS.
- Assist in activities like mounting the network infrastructure, configure IP addresses for the different equipment of laboratory-like NextSeq, Ion Torrent-PGM, 454 GS-FLX, open ports with differents protocol ssh and FTP for share files like *.sff *.fastq and flat files. Install different operative system like ubuntu, debian, centos and fedora. mount hard disck for back up information.

Publication

The draft genome sequence of Actinobacillus seminis *American Society for Microbiology*

Head Bioinfomratics *Jan. 2016 - Feb. 2016*

- Erasmo Negrete-Abascal, Fernando Montes-Garcia, Sergio Vaca-Pacheco, Abraham M. Leyto-Gil, Edgar Fragoso-Garcia, Roberto Carvente-Garcia, Sandra Perez-Agueros, Hugo G. Castelan-Sanchez, Alejandra Garcia-Molina, Tomas E. Villamar,a Patricia Sánchez-Alonso, Candelario Vazquez-Cruz. Genome Sequence of Actinobacillus seminis Strain ATCC 15768, a Reference Strain of Ovine Pathogens That Causes Infections in Reproductive Organs. doi:10.1128/genomea.01453-17.

Skills

DevOps	AWS, Docker and KVM
Programming	Shell and Python
Operative System	RedHat, Centos, Debian
Database	MongoDB
Genetic tools	Plink, Admixture, Structure, Impute, BEAGLE, Minimac,Haplogrep, SNAPPY.
Assembly tools	Fastqc,Trimmomatic, BWA, Bowtie, Newbler, Velvet, IDBA, SPADES, PAGIT, QUAST, samtools, vcftools, bedtools.
Aligment tools	BLAST, Clustal, Mugsy, Mauve.
Phylogenetic tools	PHYLIP, MrBayes, MEGA, Mesquite, FigTree, iTool.
Biobank	1k GP, HGDP, SGDP, UK-biobank.
Genetic DB	DBSNP, GWAS catalog, PharmGKB, SNPedia.
Bioinfomratics DB	NCBI, EMBL.
Languages	Spanish, English