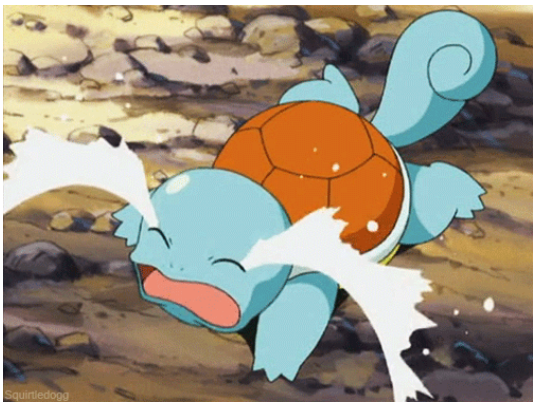
(!!!!!!!!!!!!!!!)



(!!!!!!!!!!!!!!!)

Ahora mismo no puedo acceder a mi ordenador personal debido a un problema con la batería y el cargador, y en él tengo algunos de los trabajos previos que he tenido la oportunidad de hacer usando la bioinformática. (Buena lección para que no me dé pereza meter las cosas en el disco duro…). Como la fecha límite es el 21/1, me temo que intentaré describir algo de lo que quería compartir:

Algunas cosas que me hubiese gustado adjuntar:

- plots del programa **PSMC**: es una herramienta para estimar el tamaño efectivo de una especie a lo largo de su historia evolutiva (con una resolución de miles de años) apenas a partir del genoma diploide de un solo individuo, a mí personalmente me pareció una locura. (https://github.com/lh3/psmc) .
  Utiliza la información de heterocigosidad en SNPs del genoma para estimar el Ne, y requiere de un buen filtro de calidad antes de ponerse a utilizarlo.

  Personalmente, trabajé con el genoma de varios individuos (N= 120 si no recuerdo mal) pertenecientes a una especie de pájaro en América del Norte. ¿O debería decir especie**s**? Ahí estaba el enredo… estas poblaciones habrían sufrido una radiación adaptativa hace apenas ~10000 años y el quid de la cuestión era determinar si se trataban o no realmente de la misma especie (fenotípicamente eran todos muy cucos, pero bastante diferentes). Viendo cómo evoluciona el Ne en cada una de las poblaciones, sería posible determinar en qué momento habrían divergido (o no) las unas de las otras. Además, como tiene resolución temporal, sería posible asociar estas divergencias a algún tipo de evento geológico en función de la fecha en la que se diesen.

  La ejecución era fundamentalmente desde la terminal de Linux. Primero había que pasar el genoma en formato bam por *bcftools* para hacer el pile up y un filtro de calidad de las bases teniendo en cuenta el genoma de referencia de la especie. Después se hacía el call de las posiciones de interés pasando el genoma a formato vcf.gz, y se indexaba. Todo esto para luego hacer el filtro de los SNPs (calidad mayor a 30, profundidad entre 10 y 35, …) y conseguir un archivo fasta adecuado y comprimido.

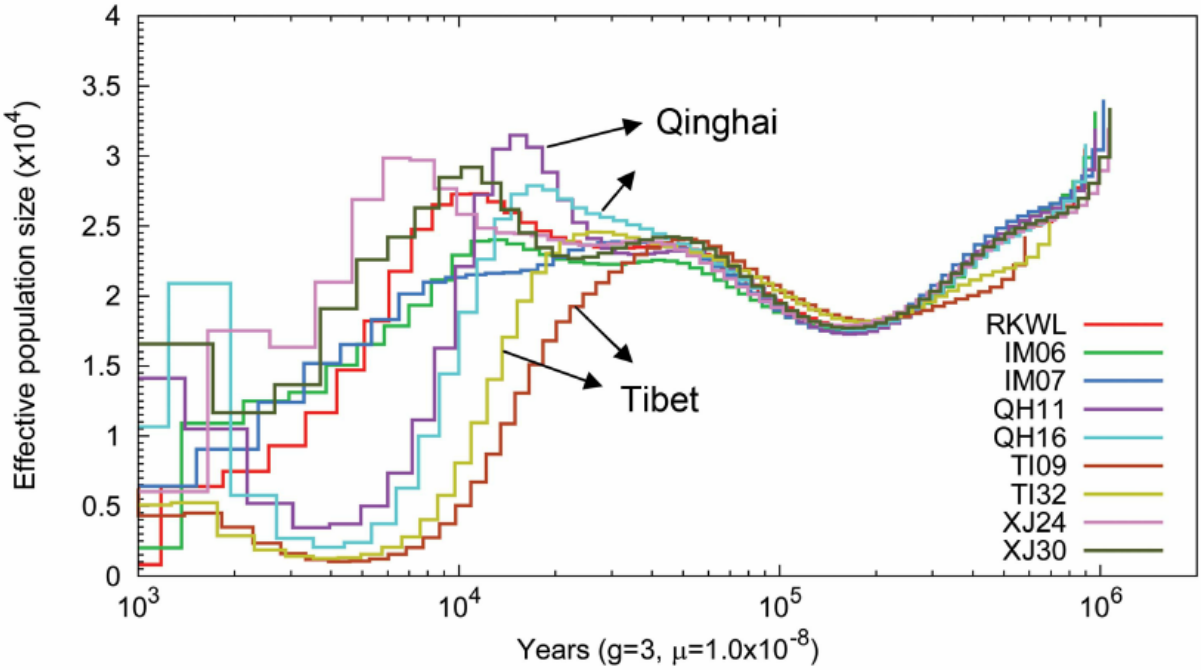
ref= /home/Documents/genome_ref.bam
bam= /home/Documents/AIK.bam

bcftools mpileup -Q30 -q30 -A -Ou -f $ref $bam
bcftools call -c -Oz -o AIK.vcf.gz
bcftools index AIK.vcf.gz
bcftools filter -i `QUAL>30 & DP>10 & DP<35` AIK.vcf.gz -Oz > AIK_filter.vcf.gz;
tabix -p vcf AIK_filter.vcf.gz; bcftools consensus -I $ref AIK.vcf.gz > AIK.fa
gzip -c AIK.fa > AIK.fq.gz

Entonces ya se podría comenzar a utilizar el programa como tal, éste usa formatos propios (psmcfa, psmc). En verdad, es muy sencillo y directo de usar: se ha de determinar el tiempo de generación, tasa de recombinación, … argumentos con los que puedes ir jugando en función de las características de tu especie. Así, se genera a un archivo .psmc gracias a un algoritmo que sigue un modelo de Markov (lo cual podría explicar sino fuese porque no tengo ni idea de cómo funciona) que se puede plotear. También admite bootstrap, aunque es computacionalmente costoso (2 semanas para 10 individuos).

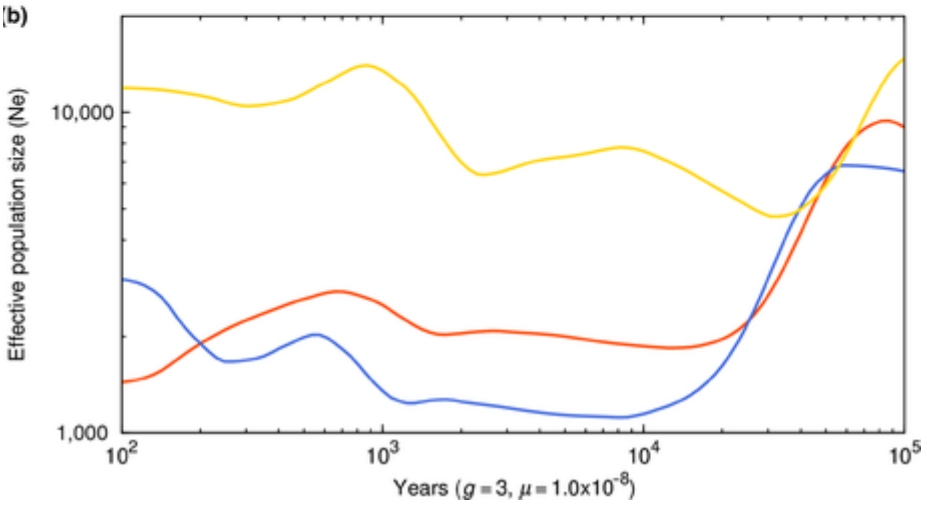
fq2psmcfa -q20 AIK.fq.gz > AIK.psmcfa
*psmc -N25 -t5 -r1 -p"4+30*2+4+6+10" -o AIK.psmc*
psmc2history.pl AIK.psmc | history2ms.pl > ms-cmd.sh
psmc_plot.pl AIK_plot AIK.psmc

He de mencionar que mis plots fueron un poco "*cuadro*" porque nuestros datos no eran apropiados para el PSMC, ya que tiene problemas con cambios de Ne recientes (~10000 años… justo a tiempo para la especiación de esos pájaros, vaya). Como resultado, tras fusionar los plots que se hacen para distintos individuos, se genera algo así:



Pairwise sequential Markovian coalescent (PSMC) analysis of nine Chinese wolf genomes reflecting the genomic distribution of heterozygous sites. Time scale on the x-axis is calculated assuming a mutation rate of 1 6 10 2 8 per generation and generation time equal to 3. doi:10.1371/journal.pgen.1004466.g004

- plots del **SMC++**: es un programa muy parecido al anterior, de hecho lo utilizamos como “segunda opinión”. (https://github.com/popgenmethods/smcpp)
  El protocolo es muy parecido al del psmc, entre ellos varía la resolución que tienen y cómo de bien funcionan para según qué especies. (*Spoiler*: tampoco funcionó bien con nuestros pájaros…).
  Los plots son algo más sencillos en comparación al psmc:



Demographic history inferred using the sequentially Markovian coalescent implemented in psmc and smc++ using all single nucleotide polymorphisms (SNPs), a generation time of 3 years and a mutation rate of 1 × 10−8. (b) Estimated histories in smc++ using the composite likelihood of four individuals from each population. The yellow line represents the site from the British Isles, the red line corresponds to western Scandinavian samples, and the blue line comprises southern Scandinavian samples. Beware the differences on the axes between the top and bottom figures, as these two methods capture variation in effective population sized through different time scales. doi: 10.1111/mec.15310

- **Hi-C:** utilizamos un montón de programas de este paquete (https://github.com/mdozmorov/HiC_tools) para analizar la estructura tridimensional del genoma de varios mamíferos. La mayor parte del tiempo no me enteraba muy bien de qué estaba haciendo el proceso, pero sí del resultado. Lo guay es que estaba escrito en Python, y para cierto problema que nos surgió muy específico de nuestros datos, pude ser capaz de modificar el código original (seguro que de manera chapucera, ¡pero funcionaba!) para que hiciese exactamente lo que nosotros queríamos.

- Plots varios en **ggplot2**: tengo varios plots representando sobre todo distribuciones de datos de variables continuas en varias especies. Principalmente, bar plots y box plots. Hacerlos me costó sudor y lágrimas, cosas que tardé milenios en encontrar cómo se hacían me fueron explicadas en los primeros 10 minutos de la P2 de TAB… ojalá haberlo hecho antes.

Siento no poder adjuntar evidencia de mis peripecias en bioinformática, ¡pero decir que esta asignatura no ha hecho más que potenciar mi interés por seguir dándole caña!