BIO 392 Bioinformatics of Genome Variations

Genomes: Core of “Personalized Health” & “Precision Medicine”

There are various names for this one concept:

Stratified, personalized, precision, individualized, P4 medicine. These are all terms used to describe notions often referred to as the future of medicine and healthcare.

This describes the use of individual genome information, concept based metadata and individually targeted therapies.

We can observe a spike in interest of P4 medicine in accordance with for ex. obama’s healthcare initiative.

Personal Genomes will soon become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

Core technologies for personalized medicine:

- Transcriptomics
- Metagenomics
- Whole genome sequencing

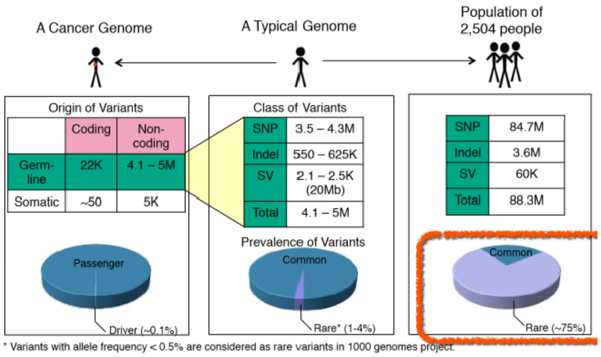
For <u>academia</u> it needs:

- Standard sample acquisition procedures & central **Biobanking**
- **Core sequencing facility**
  - Large throughput
  - Cost efficiency
  - Uniform sample
  - Data handling procedures
- Secure computing/analysis platform
- Standard **data formats** and **sample identification**
- Metadata rich, reference variant resources → **All analysis should feed into this**
- Need participation in reciprocal, international **data sharing** and **biocuration** efforts

Finding Somatic Mutations in cancer: is like finding a needle in a haystack

- Human genome (**~3 billion base pairs**) has ~**5 million variants**
- Most are rare (can only be identified as recurring when sequencing thousands of people)
- Cancer cells accumulate additional variants, only **few** of which (“**drivers**”) are relevant for the disease

<u>Graphic</u>: In the population, the rare ones add up, while the common ones are often the same between people in the population. We can see then, that most genetic variance is unknown.



Origin of Variants

	Coding	Non-coding
Germ-line	22K	4.1 – 5M
Somatic	~50	5K

Class of Variants

SNP	3.5 – 4.3M
Indel	550 – 625K
SV	2.1 – 2.5K (20Mb)
Total	4.1 – 5M

SNP	84.7M
Indel	3.6M
SV	60K
Total	88.3M

* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

Mutations and genomic rearrangements in cancer

Many cancers exhibit chromosomal rearrangements. These rearrangements can be simple, involving a single balanced fusion that preserves the proper complement of genetic information, or complex with one or more fusions that disrupt this balance. Recent technological advances have improved our ability to detect and understand these rearrangements, leading to speculation about potential causal mechanisms such as defective DNA double strand break repair and faulty DNA replication. A better understanding of these potential cancer-causing mechanisms will lead to novel therapeutic regimens to fight cancer. This review describes technological advances in methods used to detect simple and complex chromosomal rearrangements, cancers that exhibit these rearrangements, potential mechanisms for rearrangement of chromosomes, and intervention strategies designed specifically against fusion gene products and causal DNA repair/synthesis pathways.

Example: BRAF V600E (c.1799T>A) Mutation. Oncogene activation by single nucleotide alteration!

- A single nucleotide exchange Thymidine > Adenine leads to continuous RAF based activation of the MEK-ERK pathway
- BRAF V600E is a frequent mutation in >50% of malignant melanomas, but also CRC, lung ADC

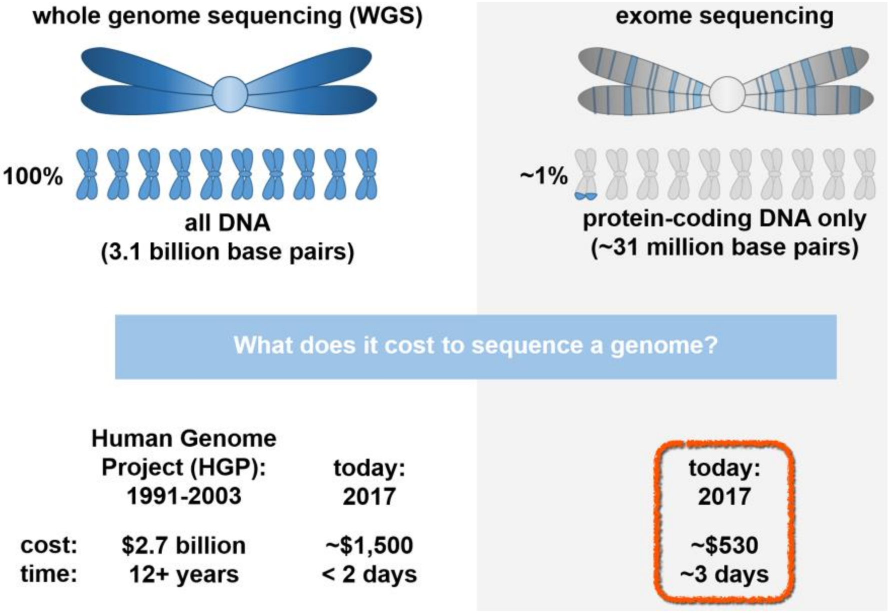
Somatic Copy Number Variations (CNVs)

A copy number variation (CNV) is when the number of copies of a particular gene varies from one individual to the next. Following the completion of the Human Genome Project, it became apparent that the genome experiences gains and losses of genetic material. The extent to which copy number variation contributes to human disease is not yet known. It has long been recognized that some cancers are associated with elevated copy numbers of particular genes.

These aren't SNP's, but large structural rearrangements that take place.

In Leukemia the most prevalent mutations are translocation mutations.

Whole Genome Sequencing (WGS) vs. **Whole Exome Sequencing (WES)**



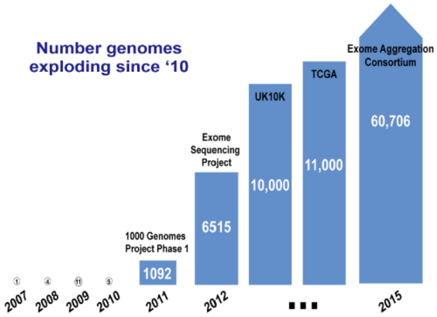
	Human Genome Project (HGP): 1991-2003	today: 2017	today: 2017
cost:	$2.7 billion	~$1,500	~$530
time:	12+ years	< 2 days	~3 days

The cost of sequencing has come a long way.

When the technology was first invented, it cost ~100 Million to sequence a human genome.

Today it costs about 1000 to sequence an entire genome!

- ➔ As a direct consequence of that, the number of sequenced genomes, is exploding.



Reference Genome Resources

UCSC Genome Browser – www.genome.ecsc.edu

- Originated from the Human Genome Project
- Most widely used general genome browser
- Many default tracks
- Many species
- Customization with “BED” files

Browser contains many tools which are developed and updated by the USCS Genome Browser Team, such as: Genome Browser, BLAT, In-Silico PCR, LiftOver

In the Genome Browser you can search for specific positions or gene symbols within a variety of species, which range from humans to rabbits to fish and many more. It is the most widely used genome browser in the world, since it has such a wide range of species to search for.

NCIB: National Institute for Biotechnology Information

- Entry point for genome reference data
- Human genome assemblies
- Human variant collections (dbVar, ClinVar, dbSNP) for downloads

ENSEMBL

- Entry point for many genome data services and collections
- Downloads (BioMart), REST API

Reference Resources for Human Genome Variants

NCBI: dbSNP

- Single nucleotide polymorphisms (SNPs) and multiple small-scale variations
- Include insertions/deletions, microsatellites, non-polymorphic variants

NCBI: dbVAR

Database of human genomic structural variation.

- Genomic structural variation
- Insertions, deletions, duplications, inversions, multi-nucleotide substitutions, mobile element insertions, translocations, complex chromosomal rearrangements

NCBI: ClinVar

- Aggregates information about genomic variation and its relationship to human health
- Database of genomic variants and the interpretation of their relevance to disease. Can tell us for a disease which variants at which gene are relevant

EMBL-EBI: EVA

- Open-access database of all types of genetic variation data from all species

Ensembl

- Portal for many things genomic

Human reference genome assemblies:

GRCh38/hg38 (Dec. 2013)
GRCh37/hg19 (Feb. 2009)
GRCh36/hg18 (Mar. 2006)

HGVS: a series of variants on one chromosome.

How we describe the data, so that it is interpretable. Or best was to compactly store this much data.

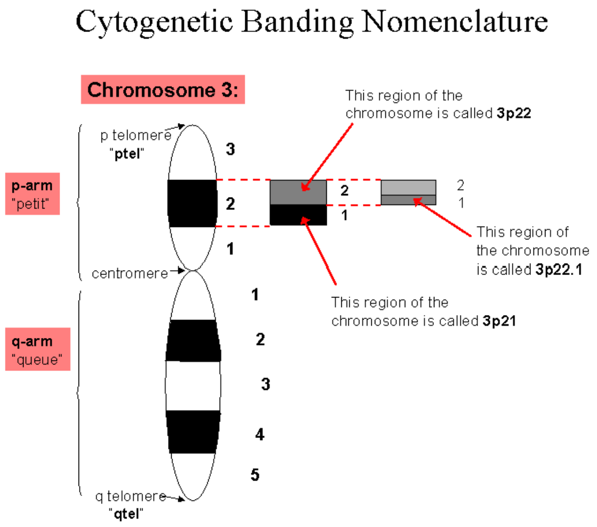
Format (one allele): **“prefix”[“change1”;”change2”]**, e.g. **g.[123G>A;345del]**

Cytogenetic banding

Short arm is p and long arm is q.

Each chromosome arm is divided into regions, or cytogenetic bands, that can be seen using a microscope and special stains. The cytogenetic bands are labeled p1, p2, p3, q1, q2, q3, etc., counting from the centromere out toward the telomeres.

At **higher resolutions**, sub-bands can be seen within the bands. The sub-bands are also numbered from the centromere out toward the telomere.


Cytogenetic Banding Nomenclature
Chromosome 3:
p telomere
"ptel"
p-arm
"petit"
3
2
1
centromere
q-arm
"queue"
1
2
3
4
5
q telomere
"qtel"
This region of the chromosome is called 3p22
2
1
2
1
This region of the chromosome is called 3p22.1
This region of the chromosome is called 3p21

"1000 genomes" (what are they, and advantages vs. problems associated with using them in genetic workflows)

- A project set out to provide a comprehensive description of common human genetic variation by applying WGS to a diverse set of individuals from multiple populations.
- Advantages: Broad representation of human genetic variation (with a much improved coverage of South Asian and African populations), use of multiple analysis strategies, increasing the quality of filtering and mapping, allowing the capture of more diverse types of genetic variants, wide availability of samples and data resulting from the project.
- Problems: Phase 1 of the 1000 Genomes Project has probably missed variants that are private, or recurrent but rare. Low-coverage sequencing, which limited CNV detection.

Genomic Variation File Formats & Tools

How big is the genome?

- 3 Billion Nucleotides (haploid) - 0.8 GB

Data science tip: Reproducibility

- Avoid manual steps of data analysis
- Keep track of all steps (using scripts)
- Use data standards
- use control version system
- save the file somewhere as a back up
- keep track of the software installed
- Don't use black boxes! ==> Open sources only
- AWK is the choice for bioinformatics

Why UNIX: Unix is for text streams

- We interpret DNA/proteins as text, UNIX is for text streams
- Data is big, Spreadsheet cannot handle them
- We need to keep track of our analysis for the sake of reproducibility ==> bash scipts
- ! Unix commands are case sensitive
- Solves the reproducible, scalability and openness for data (text) streams, but extra software might be needed
- Efficient, scalable, portable, open

<u>File formats</u>

FASTQ → FASTA

SAM → BED

- fully closed: from A to B, both included
- fully open: from A to B, both excluded
- half-open: from A to B, only A included, B excluded
- e.g. BED are format are 0 based half-open

wiggle: continuous-value data. GC%, probability scores, transcriptome data

<u>Files</u>

- Files are defined by its bytes, not the filename extension
- GRCh = genome reference consortium

Commonly used formats

- Reference genomes:

Describe the consensus DNA sequence. Multiple assemblies have been released. Usually come as FASTA

- FASTA and FASTQ (unaligned sequences)
- SAM/BED (Alignments)
- BED (Genomic ranges)
- GFF/GTF (Gene annotations)
- BEDgraphs (Genomic ranges)
- Wiggle files, BEDgraphs and BigWigs (Genomic scores)
- Indexed BEDgraphs/Wiggles

- VFG (variants)

Reference genomes: FASTA

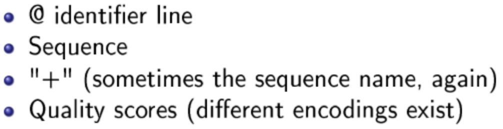
- A reference genome is a collection of contigs/scaffolds
- A contig is a stretch of DNA sequence encoded as the ltters
- Typically comes in FASTA format
- “>” line contains the scaffold name
- Following lines contain the sequence

FASTQ: short read sequencing

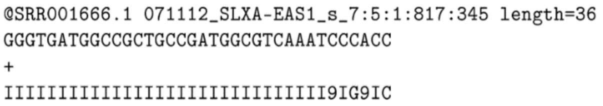
Is like a FASTA file but with a quality score. -->

First line is the identifier.

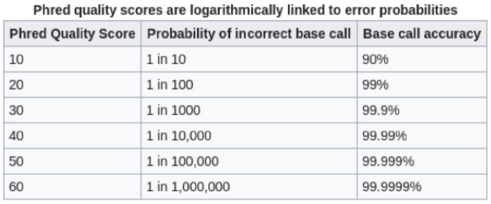
The scores are “Phred scores”: Q = -10 log10(P)



- @ identifier line
- Sequence
- "+" (sometimes the sequence name, again)
- Quality scores (different encodings exist)



```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

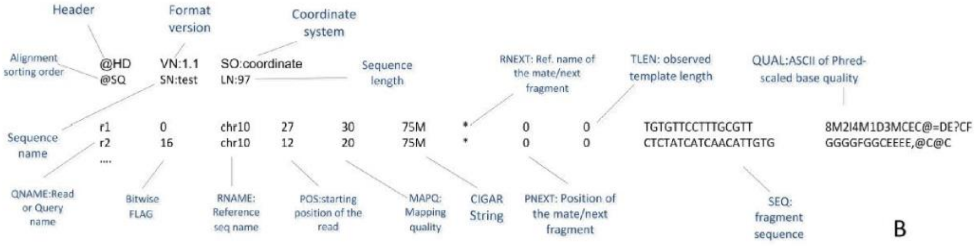


Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

SAM: Sequence Alignment Map

- Chromosome
- Locus (coordinate)
- CIGAR string



```
@HD   VN:1.1   SO:coordinate
@SQ   SN:test  LN:97

r1    0    chr10   27   30   75M   *   0   0   TGTGTTCCTTTGCGTT      8M2I4M1D3MCEC@=DE?CF
r2    16   chr10   12   20   75M   *   0   0   CTCTATCATCAACATTGTG   GGGGFGGCEEEE,@C@C
....
```

Header
Format version
Coordinate system
Alignment sorting order
Sequence length
RNEXT: Ref. name of the mate/next fragment
TLEN: observed template length
QUAL:ASCII of Phred-scaled base quality
Sequence name
QNAME:Read or Query name
Bitwise FLAG
RNAME: Reference seq name
POS:starting position of the read
MAPQ: Mapping quality
CIGAR String
PNEXT: Position of the mate/next fragment
SEQ: fragment sequence
B

SAM files are human-readable text files

BAM files are their binary equivalent

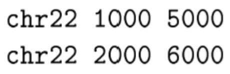
So the flow is to acquire a SAM file and the convert it to BAM.

Why? → They are smaller and easier to handle

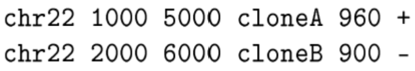
BED (Browser Extensible Data)

BED files come in different flavors

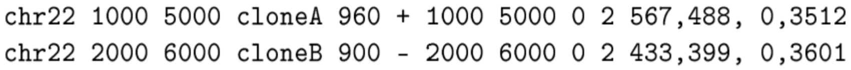
- BED3: 3 table separated columns, chromosome, start & end →
- BED6: BED3 + name, score & strand →
- BED12: →



```
chr22 1000 5000
chr22 2000 6000
```



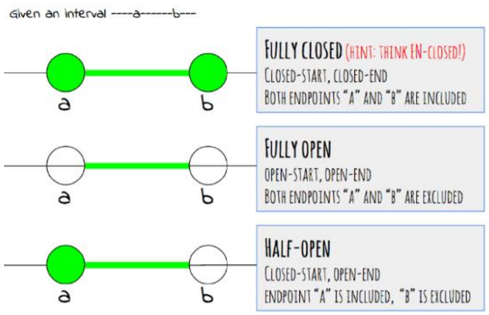
```
chr22 1000 5000 cloneA 960 +
chr22 2000 6000 cloneB 900 -
```



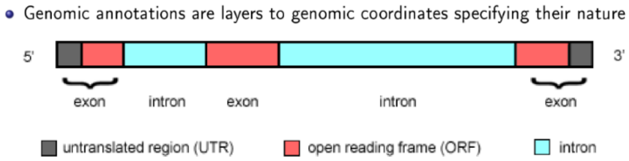
```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

<u>Counting</u>:

Depends what the base is. 0 or 1



Given an interval ----a------b---
a
b
FULLY CLOSED (HINT: THINK EN-CLOSED!)
CLOSED-START, CLOSED-END
BOTH ENDPOINTS "A" AND "B" ARE INCLUDED
a
b
FULLY OPEN
OPEN-START, OPEN-END
BOTH ENDPOINTS "A" AND "B" ARE EXCLUDED
a
b
HALF-OPEN
CLOSED-START, OPEN-END
ENDPOINT "A" IS INCLUDED, "B" IS EXCLUDED

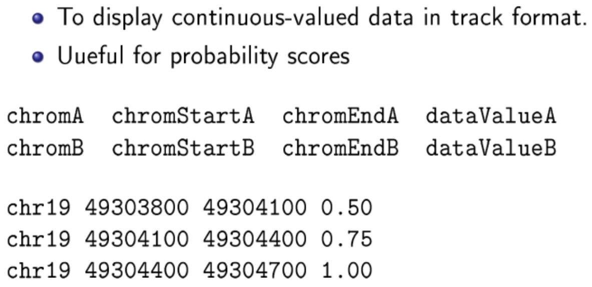


- Genomic annotations are layers to genomic coordinates specifying their nature

GFF3: General Feature Format (GFF) is a tab-delimited text file that holds information any and every feature that can be applied to a nucleic acid or protein sequence.

GTF: The Gene transfer format (GTF) is a file format used to hold information about gene structure. It is a tab-delimited text format based on the general feature format (GFF), but contains some additional conventions specific to gene information. A significant feature of the GTF that can be validated: given a sequence and a GTF file, one can check that the format is correct. This significantly reduces problems with the interchange of data between groups.

BEDgraph: →



- To display continuous-valued data in track format.
- Ueful for probability scores

```
chromA  chromStartA  chromEndA  dataValueA
chromB  chromStartB  chromEndB  dataValueB

chr19 49303800 49304100 0.50
chr19 49304100 49304400 0.75
chr19 49304400 49304700 1.00
```

Wig files:

- To display continuous data
- GC percent, probability scores, and transcriptome data.
- Data is not sparse. Wiggle data elements must be equally sized (Step)

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

```
variableStep chrom=chr2 span=5
300701 12.5
```

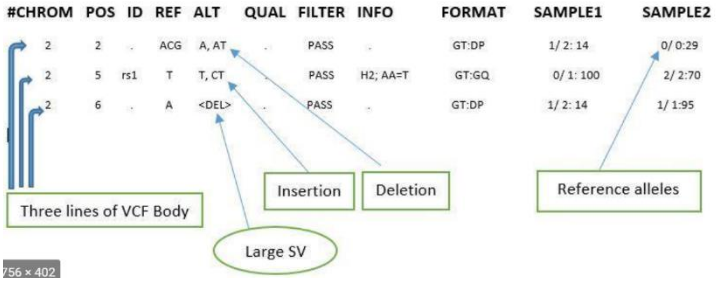
Same things but much higher information density

How to represent the scores of 300 nucleotides from chr3, starting from 400601, where the first 100 nt are scored 11, the next 100 nt 22, and the last 33? →

```
fixedStep chrom=chr3 start=400601 step=100 span=5
11
22
33
```

VCF (Variant Call Format)

- Standard file format for storing variation data
- Unambiguous, scaleable, and flexible
- 8 columns: CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO



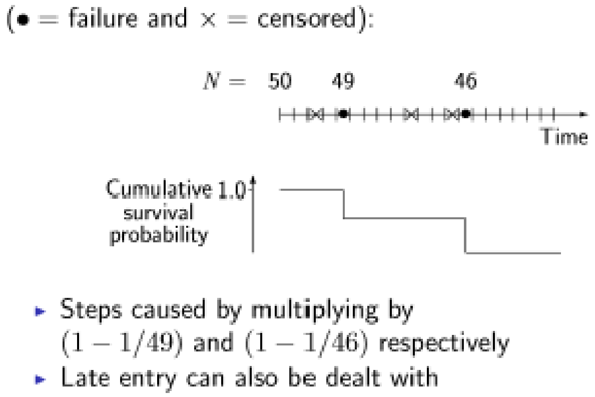
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
2	2	.	ACG	A, AT	.	PASS	.	GT:DP	1/ 2: 14	0/ 0:29
2	5	rs1	T	T, CT	.	PASS	H2; AA=T	GT:GQ	0/ 1: 100	2/ 2:70
2	6	.	A	<DEL>	.	PASS	.	GT:DP	1/ 2: 14	1/ 1:95

File formats sorted by costs:

SAM > BAM > FASTA > VCF

Survival Analysis

Kaplan-Meyer Plot


(● = failure and × = censored):

- Steps caused by multiplying by $(1 - 1/49)$ and $(1 - 1/46)$ respectively
- Late entry can also be dealt with

Kaplan-Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. In clinical trials or community trials, the effect of an intervention is assessed by measuring the number of subjects survived or saved after that intervention over a period of time. The time starting from a defined point to the occurrence of a given event, for example death is called as survival time and the analysis of group data as survival analysis. This can be affected by subjects under study that are uncooperative and refused to be remained in the study or when some of the subjects may not experience the event or death before the end of the study, although they would have experienced or died if observation continued, or we lose touch with them midway in the study. We label these situations as censored observations. The Kaplan-Meier estimate is the simplest way of computing the survival over time in spite of all these difficulties associated with subjects or situations. The survival curve can be created assuming various situations. It involves computing of probabilities of occurrence of event at a certain point of time and multiplying these successive probabilities by any earlier computed probabilities to get the final estimate. This can be calculated for two groups of subjects, also their statistical difference in the survivals. This can be used in Ayurveda research when they are comparing two drugs and looking for survival of subjects.

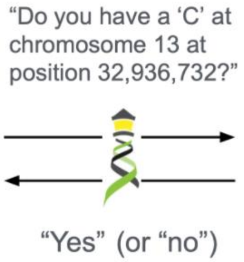
Censored = an ongoing study, some have been in it for long and some have just started the treatment. It is observation after treatment given.
The next interval then is calculated with a different value.

PLINK: the five main domains of function:

- data management
- summary statistics
- population stratification
- association analysis
- identity-by-descent estimation.

Genomic Privacy: genome "<u>beacons</u>"

Any hospital or research entity can choose to 'beaconize' their omics dataset without compromising the privacy or the ownership of the dataset. Therefore, helping the worldwide community of researchers and assisting science through the power of data. One can query the database like this: →



"Do you have a 'C' at chromosome 13 at position 32,936,732?"
"Yes" (or "no")