Day 1

- HPC setup (access to gaia)
- Bash continued
- Lecture: Introducing Family Genomics
- Start tutorial on gaia

Day 2

- Run tutorial on gaia
- Visualize sequencing data using IGV

Demo :

Analysing Whole Exon Sequencing Data of a Family

Dr. Patrick May

Head of Genome Analysis/Bioinformatics Core (LCSB/Luxembourg)
Family Genomics Group (ISB/Seattle)

patrick.may@uni.lu

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Outline

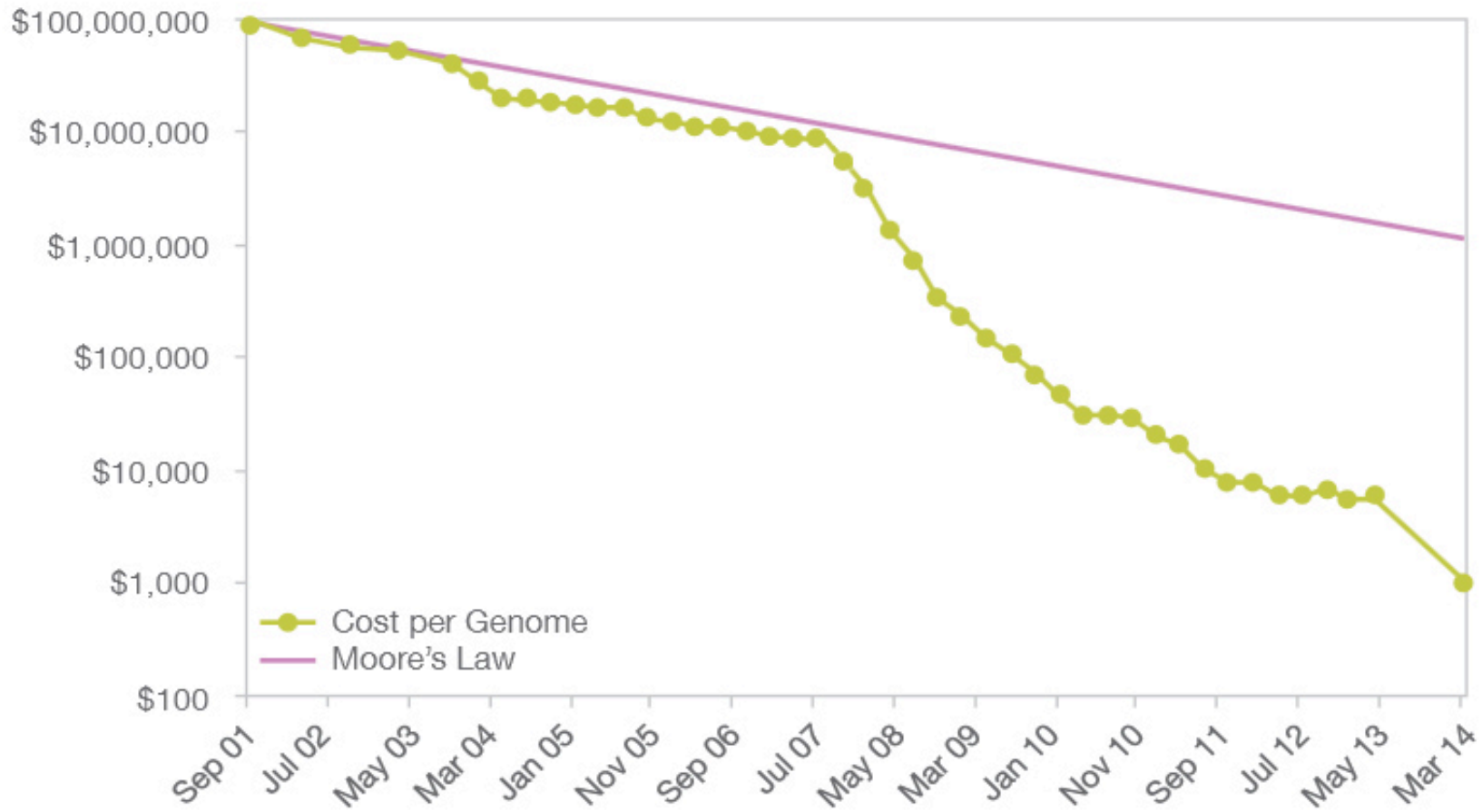
- Family Whole Exome Sequencing (WES)
- WES analysis pipeline
- GATK best practices
- Variant analysis pipeline
- Demo: Corpasome
  - Getting started on Gaia (hpc.uni.lu)

Family Whole Exome Sequencing

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Sequencing costs



LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

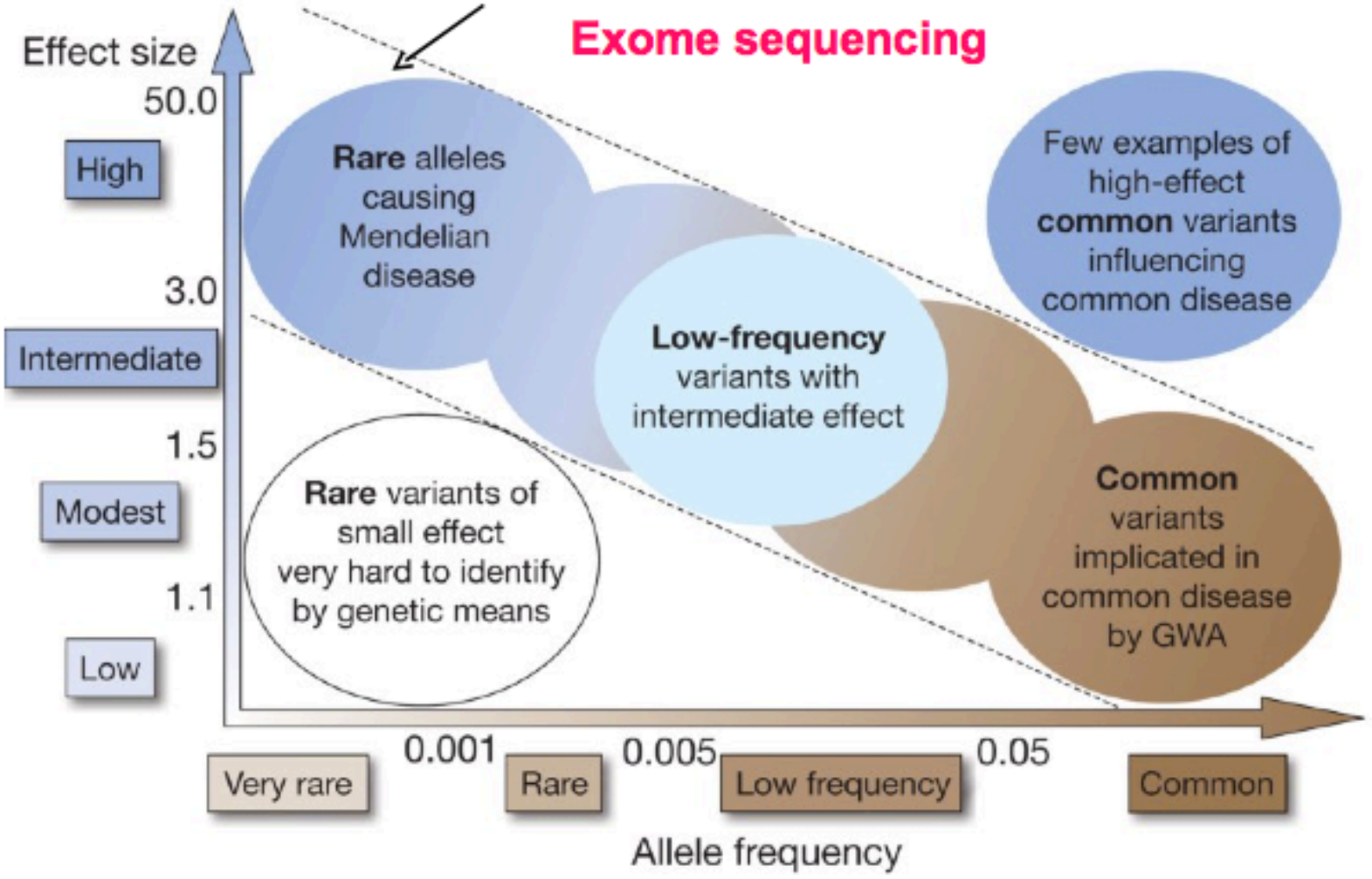
Sequencing technologies

- Illumina dominates the market (HiSeq, NextSeq, Xten)
- There are others, such as CGI, recently acquired by BGI.
- Technologies will change
  - but, general principles taught in this course unlikely to change much
- Don't automatically assume that sequencing data is perfect – understand the "error characteristics" of the technology you are using.
  - understand "quality score" thresholds for your technology
- WES data is problematic
  - coverage tends to tail off near the edges of exons
- WGS data is problematic, too
  - coverage can vary, biases can be subtle

Why whole exon sequencing ?

- Exome makes 2% of the human genome, but contains ~85% of known disease-causing variants
- WES is a cost-effective alternative to whole-genome sequencing (WGS)
- In case/control (GWAS) study only common variants can be detected and only with large sample sizes
- WES allows for the detection of rare variants

Rare and common variants



Manolio TA, et al (2009). Finding the missing heritability of complex diseases. Nature., 461(7265), 747

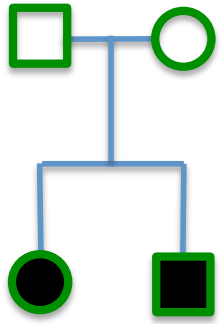
NGS Genetics

- Mendelian diseases
  - One or few pedigrees with one or more affecteds
    - Past: linkage analysis
    - Today: WES/WGS of all individuals
- Complex diseases:
  - Groups of cases and controls
  - Many families with one mode of inheritance
    - Past: GWAS (Genome wide association studies)
    - Today: WGS/WES or targed resequences

Why family sequencing ?

- Next-generation sequencing allows the cost effective characterization of whole exomes/genomes
- In case/control (GWAS) studies only common variants can be detected with large sample sizes
- Families allow for the detection of *de novo* mutations
- Inheritance pattern for both rare and common alleles can be analyzed across the entire genome
- Eliminate false positives: Mendelian inheritance errors
- Reconstruct no-calls using pedigree
  - in theory, can infer (or 'impute') missing data even in individuals not sequenced, or who have been genotyped with a modern genotyping panel.

The Evolution of Family Genomics

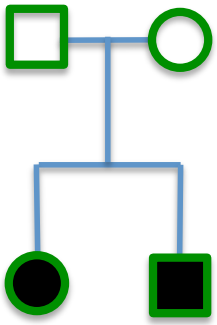


Miller-Syndrome
Roach et al. 2010

LCSB UNIVERSITÉ DU LUXEMBOURG

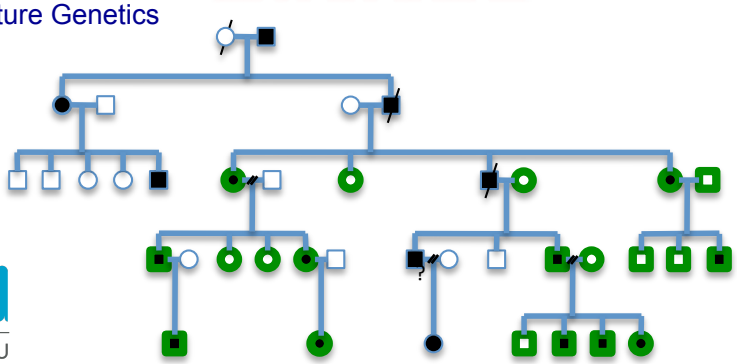
Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

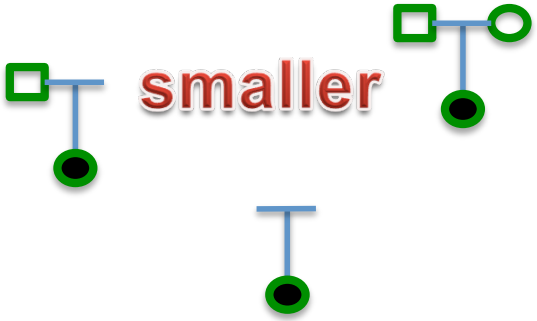
The Evolution of Family Genomics



Miller-Syndrome
Roach et al. 2010, Science

Generalized epilepsies with febrile seizures plus
Schubert et al. 2014, Nature Genetics



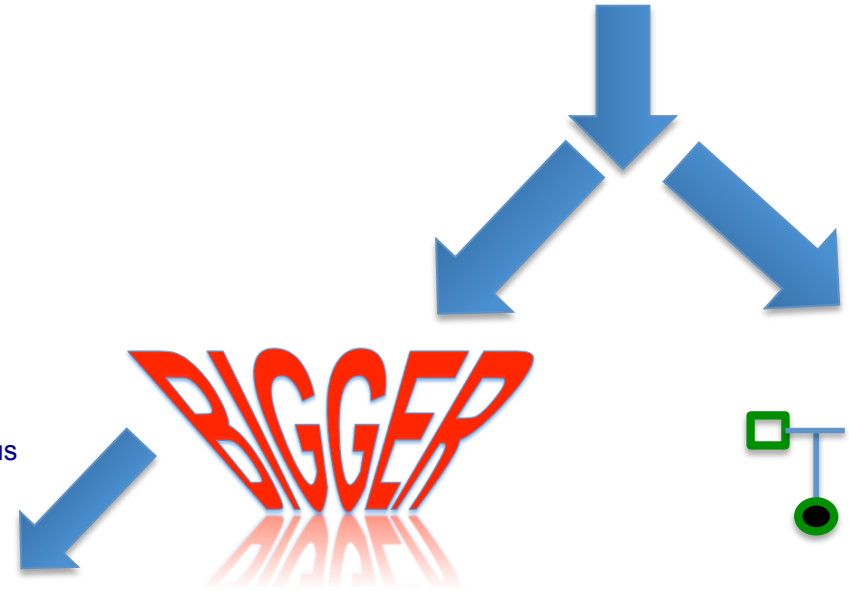


Epileptic encephehalopathies
Nava et al. 2014, Nature Genetics
EuroEPINOMICS, Epi4K et al, 2014, AJHG

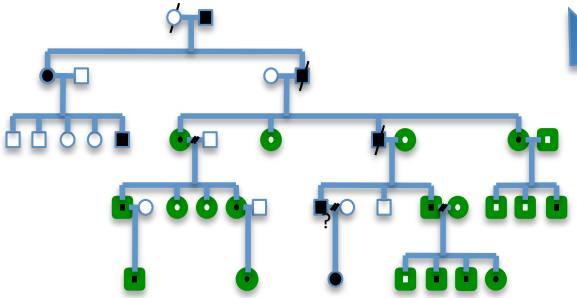
LCSB UNIVERSITÉ DU LUXEMBOURG

EUROCORES Programme
EuroEPINOMICS
Functional Genomic Variation in the Epilepsies

The Evolution of Family Genomics

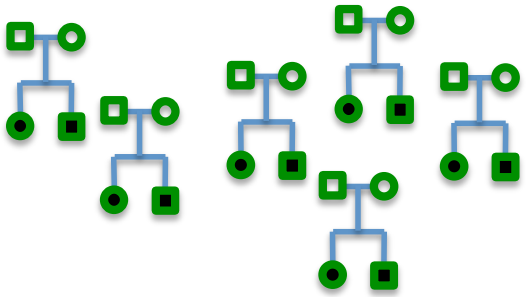


Generalized epilepsies with febrile seizures plus
Schubert et al. 2014, Nature Genetics

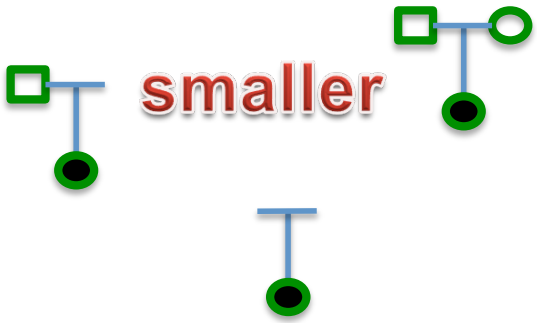


Bigger pedigrees

More pedigrees



Autosomal Recessive Rare Epilepsy Syndromes
Hardies, May, et al, submitted
EuroEPINOMICS RES consortium



Epileptic encephehalopathies
Nava et al. 2014, Nature Genetics
EuroEPINOMICS, Epi4K et al, 2014, AJHG

LCSB UNIVERSITÉ DU LUXEMBOURG

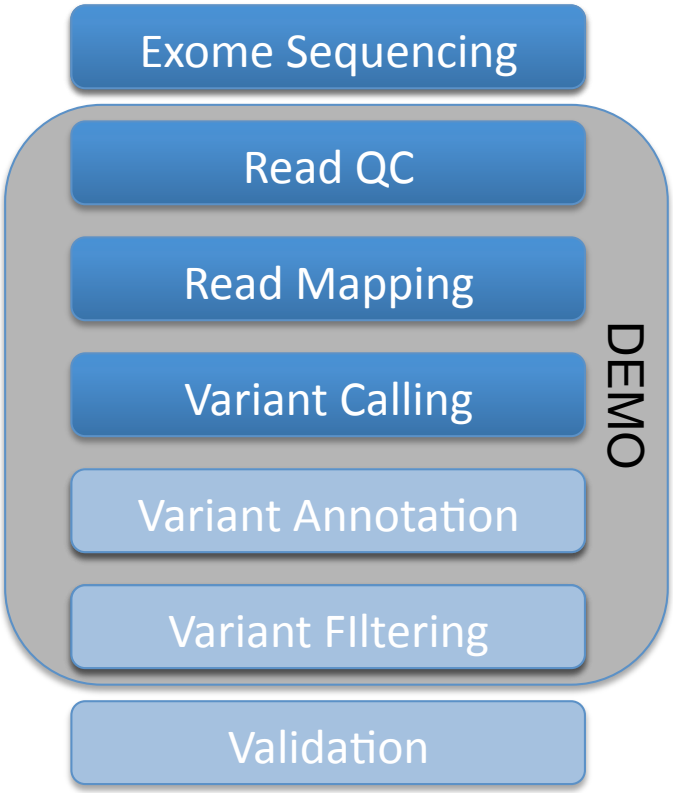
EUROCORES Programme
EuroEPINOMICS
Functional Genomic Variation in the Epilepsies

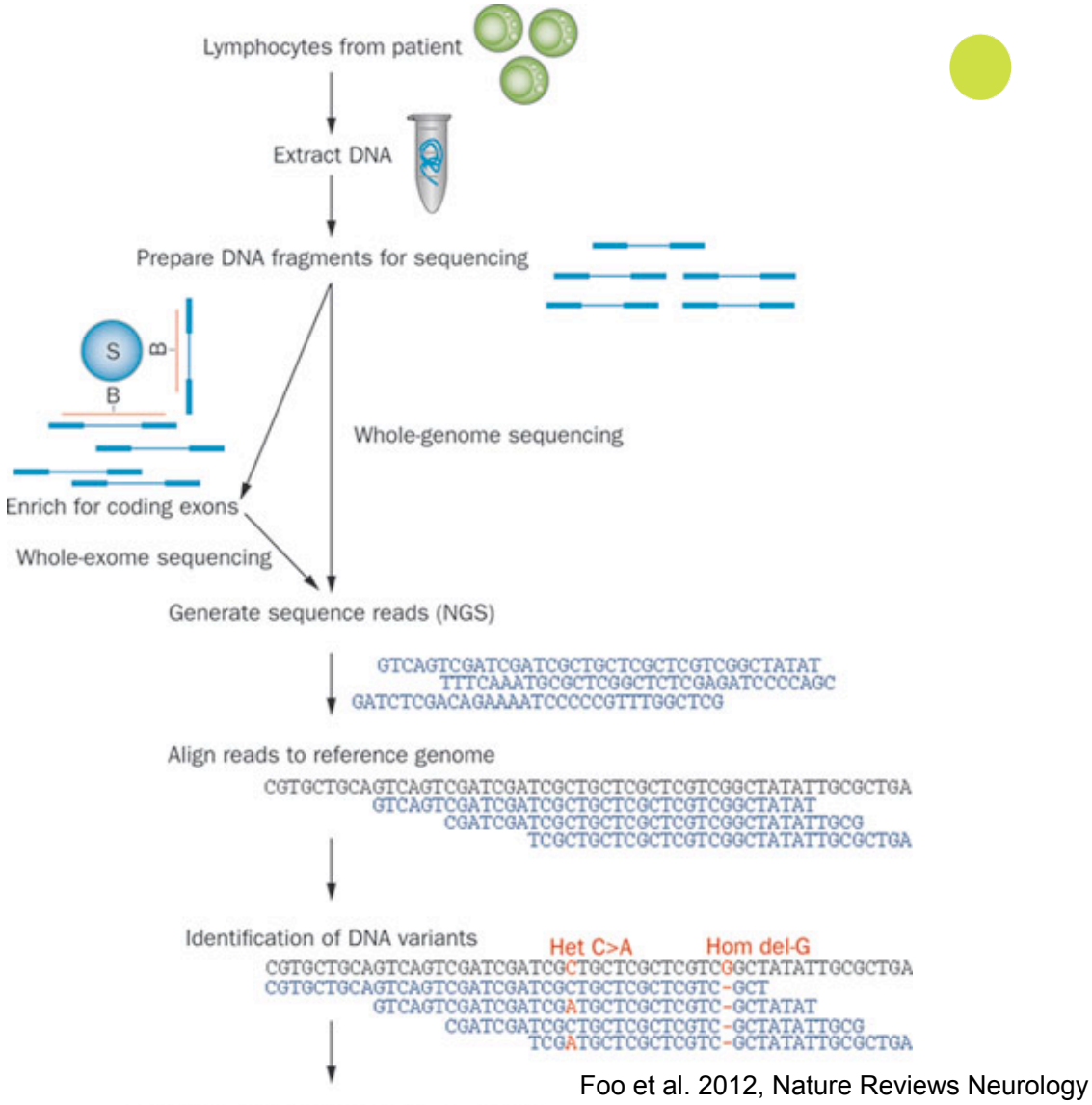
WES analysis pipeline

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

WES analysis





Foo et al. 2012, Nature Reviews Neurology

LCSB UNIVERSITÉ DU LUXEMBOURG

The FASTQ format

The FASTQ format (Illumina)

```
1.  @FCC189PACXX:2:1314:19975:86201/2
2.  CTCTCTTTCTCTCTTTCTCTCTTTCTTTTCTTTTTTCTT ...
3.  +
4.  bb_eeeeegggfghhhiagfihhhhiiiihiiiiiiafg ...
```

1. Instrument name, flowcell id, coordinates within the tile, first or second read pair
2. The sequence of the read
3. Optional description
4. Quality values in ASCII (33 + Phred scaled Q)

Sequences from NGS machines are stored in that format!

LCSB UNIVERSITÉ DU LUXEMBOURG

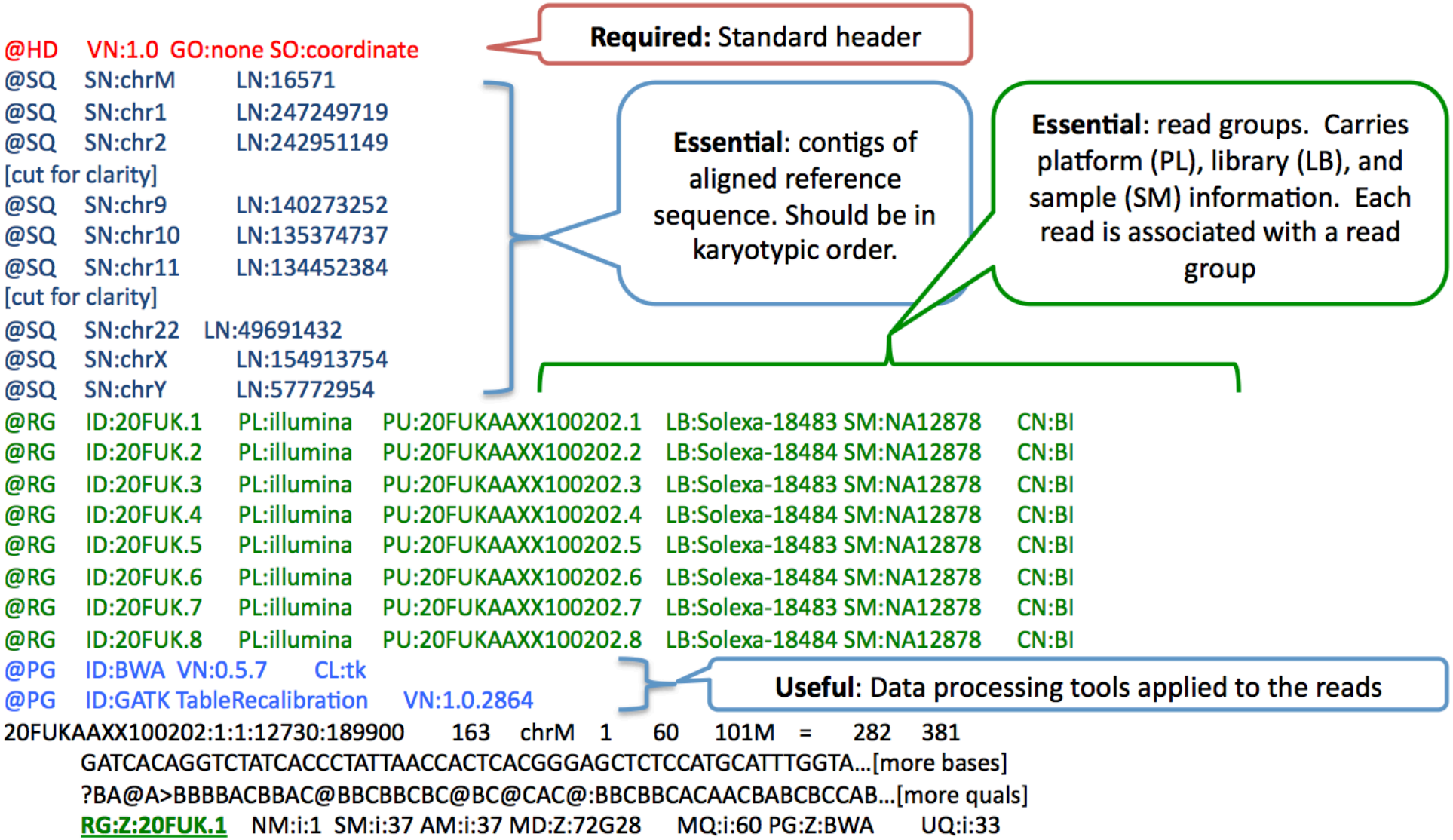
https://en.wikipedia.org/wiki/FASTQ_format

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

The SAM alignment format

- http://samtools.github.io/hts-specs/SAMv1.pdf
- Has become the standard for storing NGS alignment data
- BAM format is the binary compressed version of SAM including indexing capabilities for fast access
- Many tools support this format
- Developed at Welcome Trust Sanger Institute by Heng Li and published in 2009 (Bioinformatics)

BAM headers: an essential part of a BAM file



The SAM alignment format

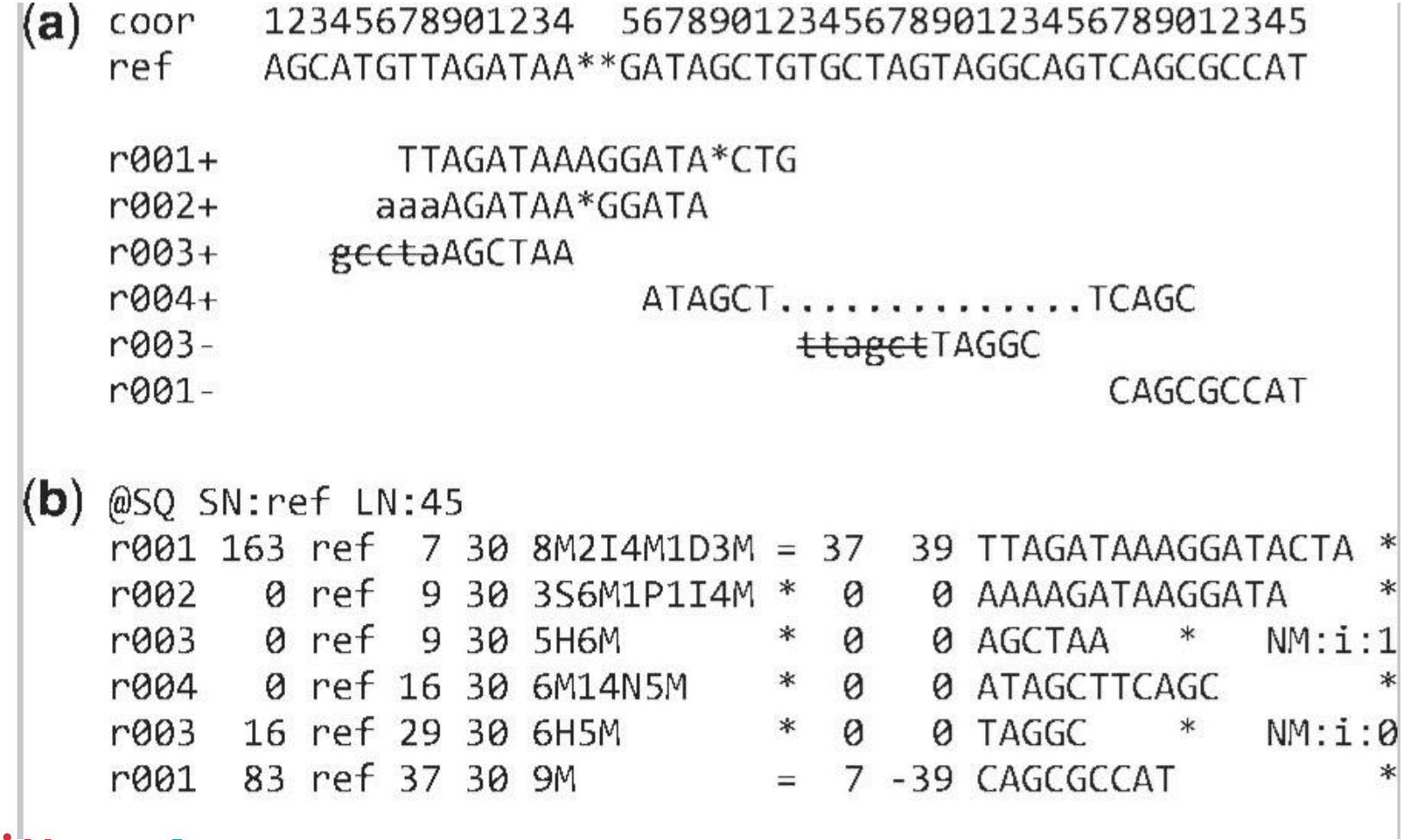
No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

The SAM alignment format

The CIGAR string

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

The SAM alignment format



(a)

```
coor    12345678901234  5678901234567890123456789012345
ref     AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

r001+         TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     gcctaAGCTAA
r004+                   ATAGCT..............TCAGC
r003-                        ttagctTAGGC
r001-                                    CAGCGCCAT
```

(b)

```
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Short read alignment

- Before NGS:
  - FASTA, BLAST, MEGABLAST, SSAH2, BLAT
- NGS produces short reads in high throughput, therefore new specialized fast aligners were needed

Short read aligner

- ELAND, RMAP, MAQ, ZOOM, SEQMAP, CLOUDBURST, SHRIMP:
  - hashing reads and scan reference, flexible memory footprint, high overhead when scanning few reads
- SOAPV1, PASS, MOM, PROBEMATCH, NOVOALIGN, RESEQ, MOSAIK, BFAST:
  - hash the genome, parallelizable, require large memory to build reference index, speed sensitive to sequence errors
- SOAPV2, BOWTIE, **BWA**:
  - Burrows-Wheeler Transform (BWT), prefix tree with small memory footprint

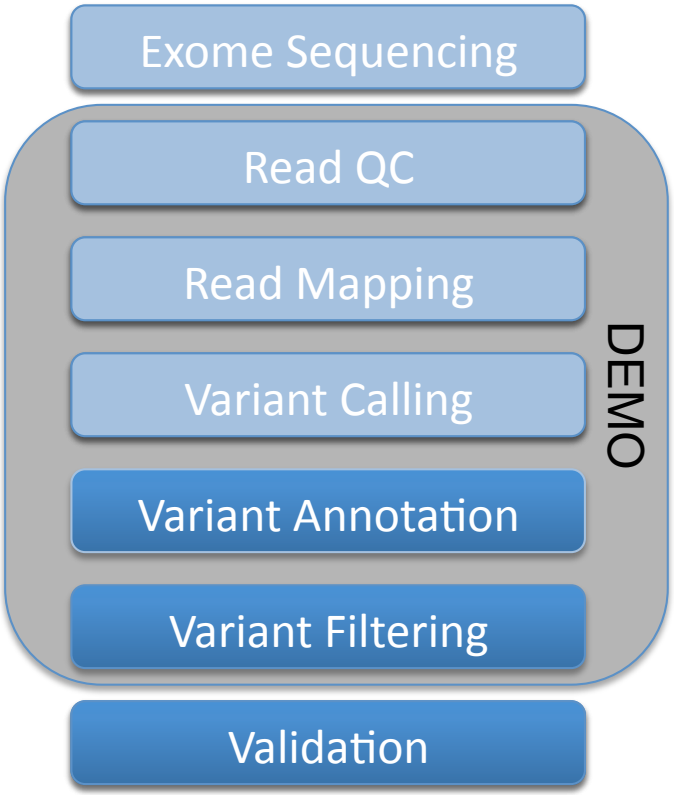
BWA

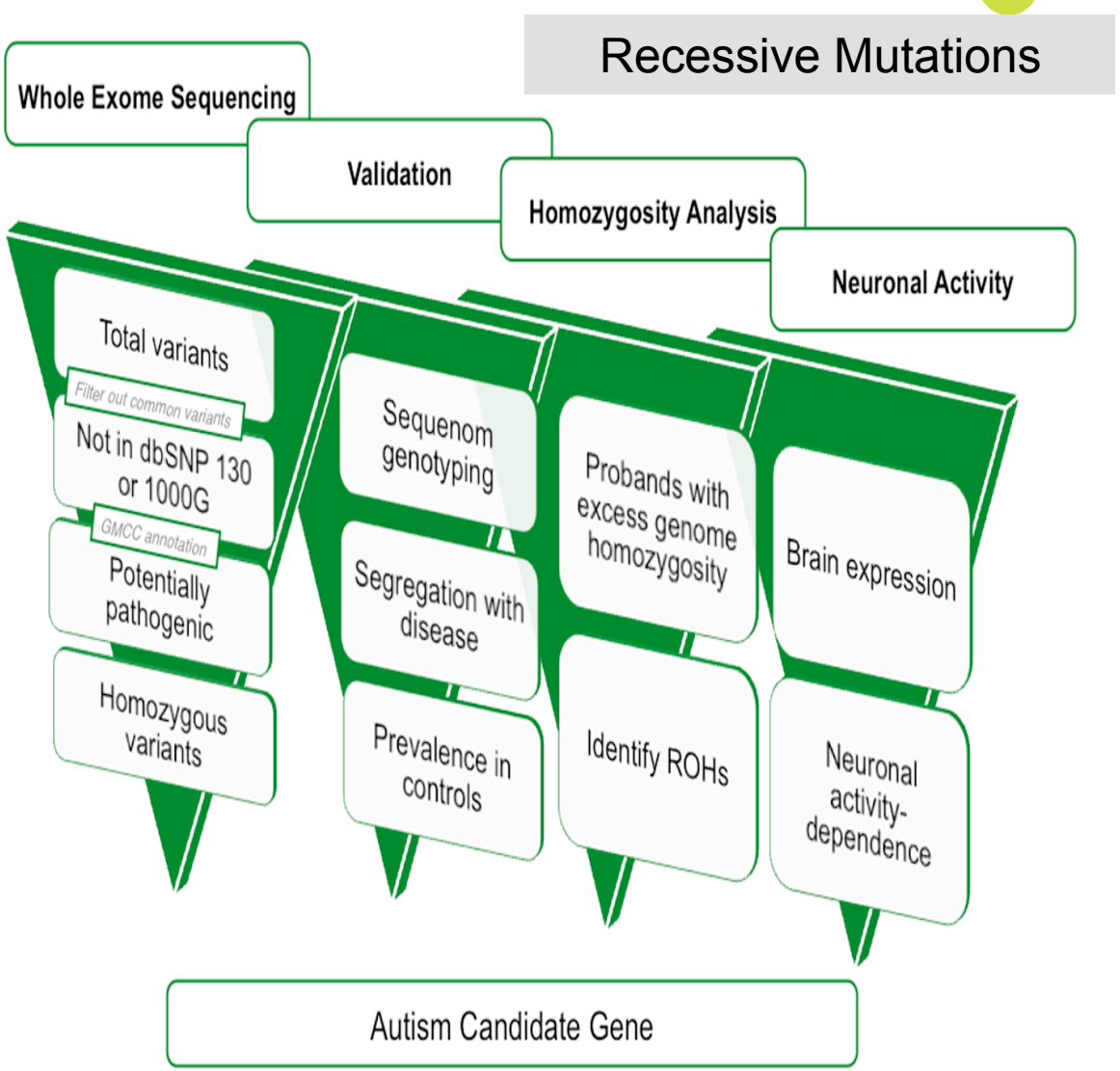
- Uses the Burrows-Wheeler transform algorithm
- Fast and moderate memory footprint
- For different platforms, SE or PE
- Gapped alignments
- Non-unique reads are placed randomly with a mapping quality=0
- Output alignments in SAM format
- Li, H. and Durbin, R., Fast and accurate short read alignment with Burrows- Wheeler transform. *Bioinformatics* **25** (14), 1754 (2009)

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

WES analysis





Charour et al. 2012, PLOS Genetics

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

GATK

Best Practices

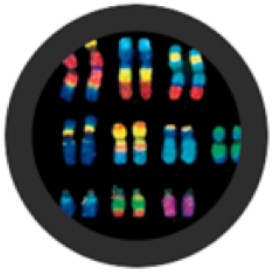
http://www.broadinstitute.org/gatk/guide/best-practices

DNAseq

About
Introduction to GATK
Citing GATK
GATK is cited by...
Real world impacts
Development team

GATK Development Team

The Genome Sequencing and Analysis Group @ Broad



The **Genome Sequencing and Analysis Group** (GSA) in Medical and Population Genetics at the Broad Institute is a team of computational biologists, software engineers, and hosted students and researchers developing algorithms for next generation DNA sequencers for medical and population genetics and cancer applications, as well as applying these algorithms to answer fundamental scientific questions.

GSA has extensive experience with processing of next-generation DNA sequencer data as well as genotyping and validation data along with downstream analysis of this data for medical and population genetics studies. The method development arm of GSA has created a powerful framework in the The Genome Analysis Toolkit for analysis of next-generation sequencing data and analysis of variation discovered by NGS. These tools are now widely used in many NGS projects, including the 1000 Genomes Project, The Cancer Genome Atlas, the Broad's production sequencing pipeline, as well as at many other sequencing centers and individual labs with sequencing machines.

GROUP MANAGER

Eric Banks

METHODS DEVELOPMENT

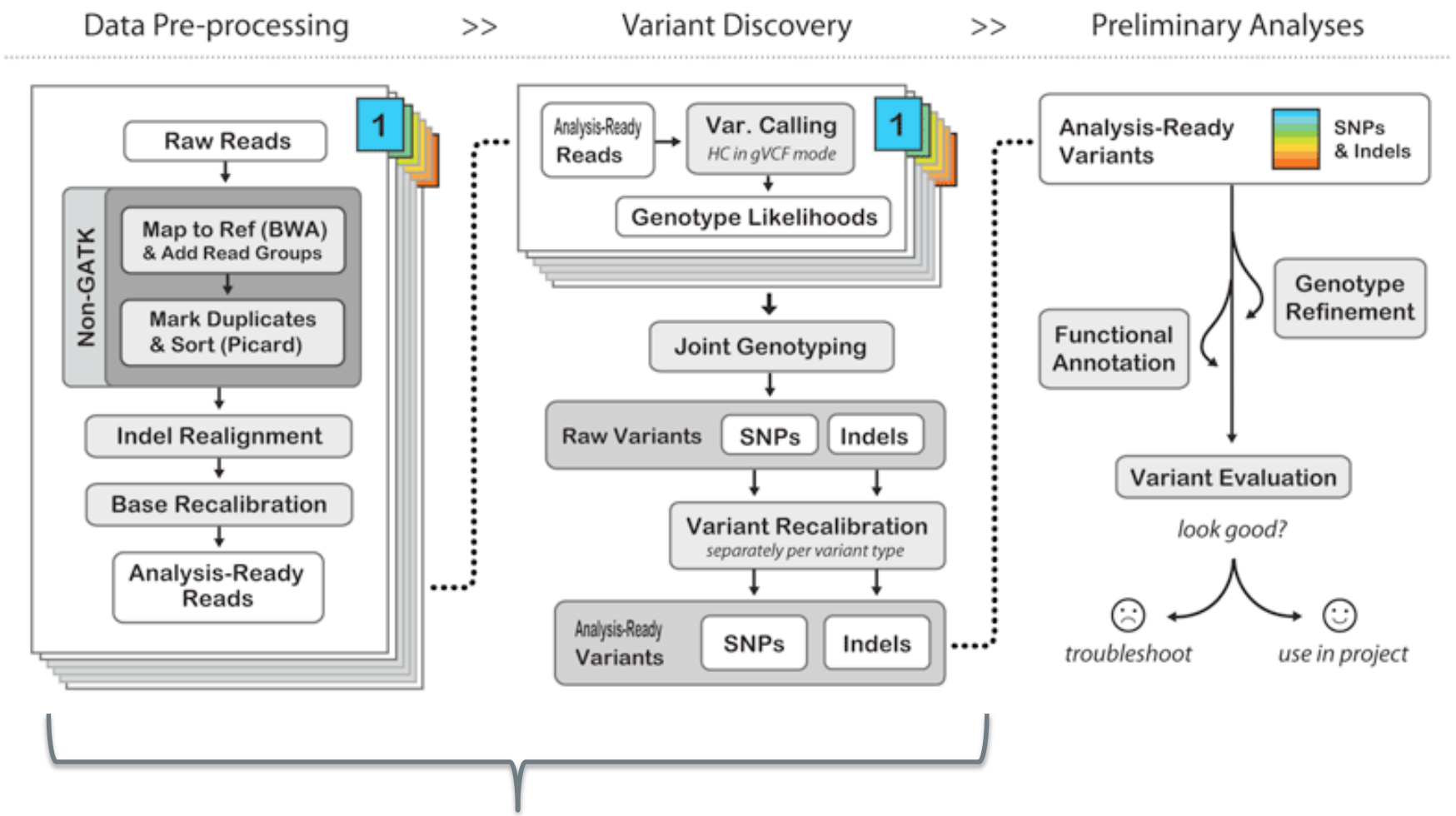
Ryan Poplin, Team Lead

Ami Levy-Moonshine

Eric Banks, PhD, is the Group Leader of the Methods Development Team in the Genome Sequencing and Analysis Group at the Broad Institute. His team develops many of the highly used analysis tools that make the GATK so powerful.

Before working at the Broad, Eric completed his PhD in Computational Biology under Professor Mona Singh at Princeton University, where he held Gordon Wu and PICASso fellowships. The title of his thesis was Algorithms for Analyzing and Interrogating Protein Interaction Networks. Eric worked both as an undergraduate and for his Masters in Engineering at MIT with Professor Bonnie Berger.

GATK Best Practices - DNAseq



DEMO

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Removal of duplicated reads

- Introduced during library creation/amplification
- Optical duplicates
- Many duplicates of a read with wrong indel can mask the correct one
- Will result in high read depth and can be the cause of false positives
- Duplicates: Identical 5’ coordinates and orientations
- Best: Read pair having highest sum of base qualities
- Can be removed with samtools or picard

Base Quality Score Recalibration (BQSR)

- Observed error rates differ from raw base quality scores
- More over, the base quality is not evenly distributed in a read: machine cycle bias, sequence context, sequencing chemistry effects
- BQSR is:
  - the sum of the global difference between reported quality scores and the empirical quality
  - plus the quality bin specific shift
  - plus the cycle x qual and dinucleotide x qual effect
- Sites of known variations are taken into account

Local realignment of reads

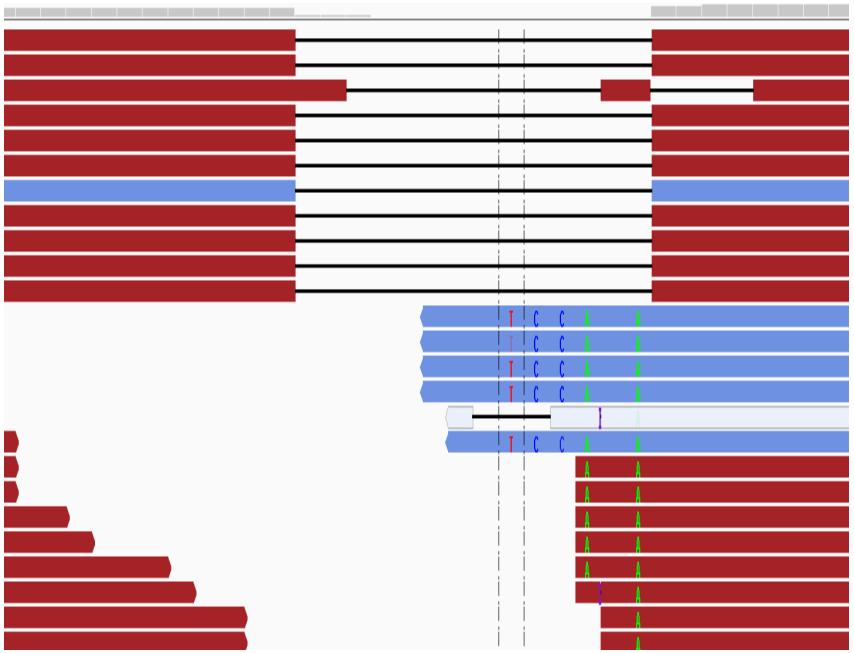


Figure 7: (a) Before Realignment

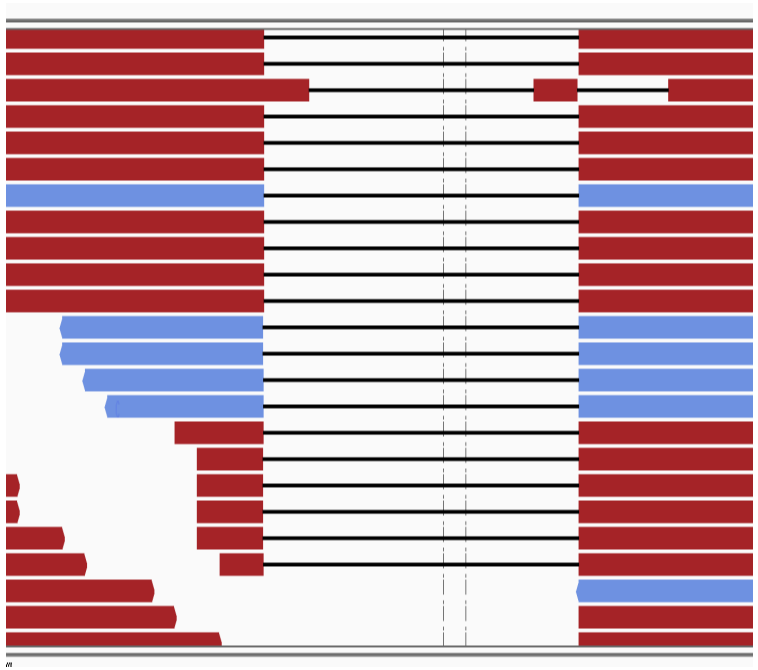


Figure 7: (b) After Realignment

Variant analysis pipeline

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Variant calling - Samtools

- mpileup is used to generate a pileup of all bases at a certain position
- each line represents a genomic position, consisting of chromosome, coordinate, reference base, read bases, read qualities and alignment mapping qualities

Institute for Biology
Enhancing Life.

Variant calling - Samtools

- Bcftools (part of samtools package) is used to convert between VCF (variant call format) and BCF (binary VCF), and to call variants
- mpileup output in BCF format can directly piped into bcftools

General VCF format

SAM/BAM + related specifications: https://github.com/samtools/hts-specs

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002
20  14370     rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|
    0:48:1:51,51 1|0:48:8:51,51
20  17330     .         T      A       3    q10    NS=3;DP=11;AF=0.017
          GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3
20  1110696                rs6040355  A      G,T    67   PASS   NS=2;DP=10;AF=0.333,0.667
          GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2
20  1230237                .          T      .      47   PASS   NS=3;DP=13;AA=T
          GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51
20  1234567                microsat1  GTC    G,GTCT 50   PASS   NS=3;DP=9;AA=G       GT:GQ:DP    0/1:35:4
          0/2:17:2
```

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Where is the genotype?

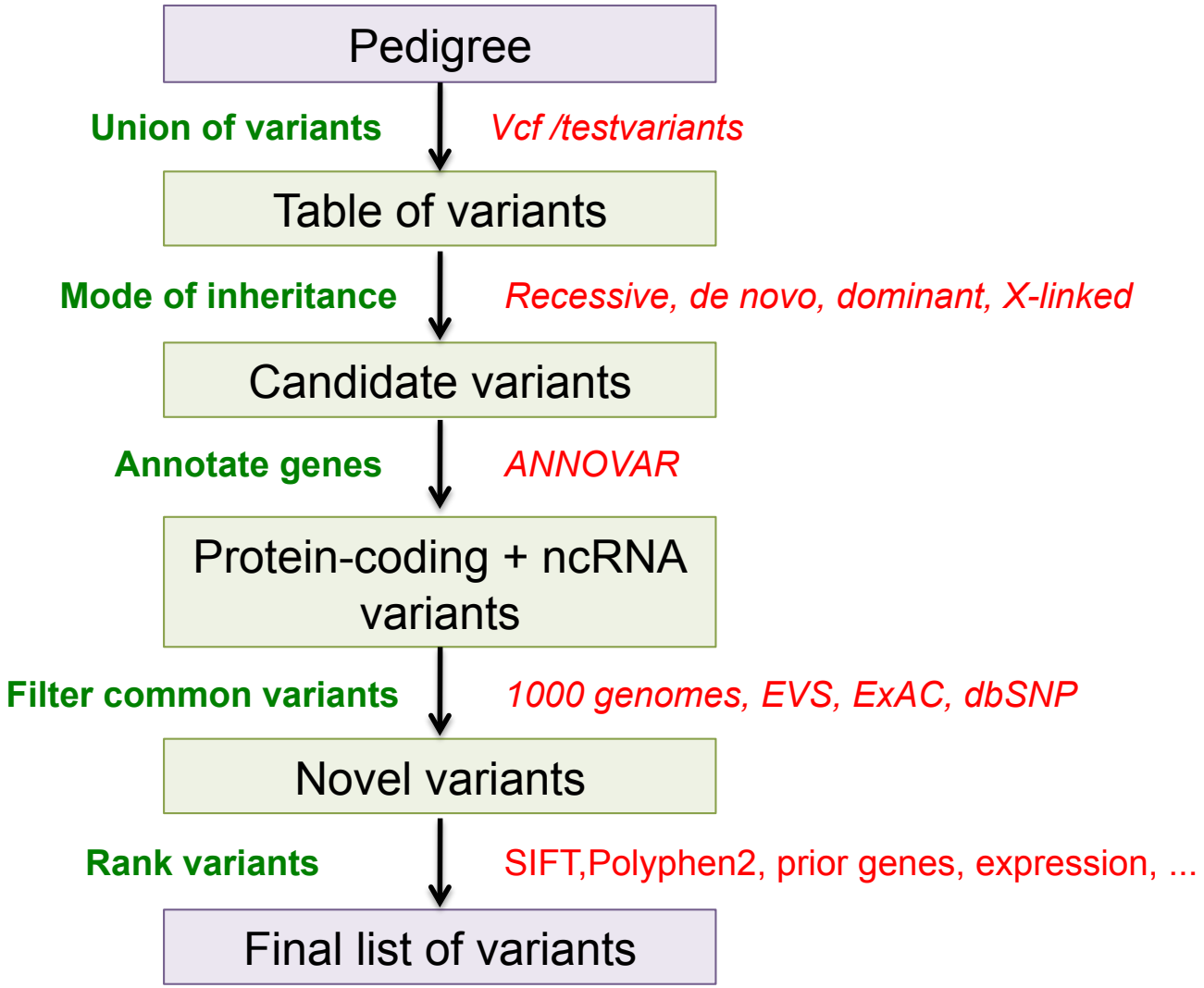
- The genotype is decoded in the PL format-tag
- eg: ref=C; alt=A,G; PL=7,0,37,13,40,49
- PL is a list of phred-scaled genotype likelihoods
- From the the given example, the most probable genotype is C/A ($10^0=1$)

GT:	CC	CA	AA	CG	AG	GG
PL:	7	0	37	13	40	49
=:	$10^{-0.7}$	10^{0}	$10^{-3.7}$	$10^{-1.3}$	10^{-4}	$10^{-4.9}$

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Variant Analysis Pipeline

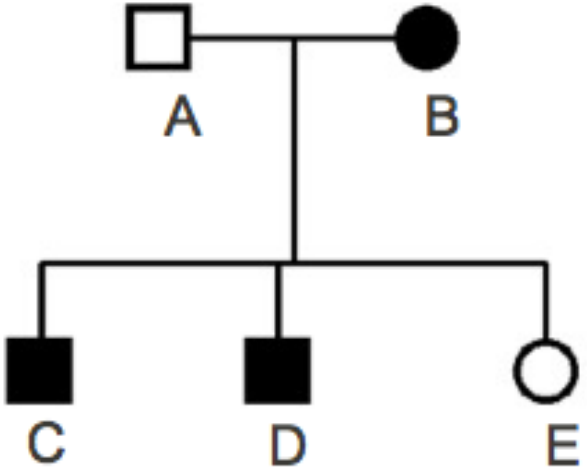


LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Mode of inheritance

- Diseases can be dominant, recessive or X-linked
  - compound heterozygosity (recessive)
  - *de novo* variants (trio)
  - incomplete penetrance
  - time of disease onset (modifier)



Dominant: heterozygous in **B**,**C**,**D**
reference in A,E

Recessive: homozygous in **B**,**C**,**D**
heterozygous in A,E

Annotation: ANNOVAR

(Wang et al. NAR 2010)

- Annotation of exonic, intronic, UTR, splice-site, up-/ downstream and ncRNA variants
- Different annotation sets are used
  - *refGene, ucsc, ensembl, ccds, gencode*
  - non-coding tracks: *ucsc, encode*
- All annotations are combined per variant
- Filtering according to pre-compiled variant sets, e.g., *EVS, 1000G, CG69*
- Searching for variants previously annotated in *OMIM, ClinVar, HGMD, GWAS, dbGAP* or *dbSNP*

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

(Lazy) Filtering: False Positives

```
next if ($gene =~ /^OR\d+\D+\d+P?/);
#skip olfactory receptors

next if ($gene =~ /^MUC\d+P?/);
#skip MUC genes

next if ($gene =~ /^HLA/);
#skip HLA genes

@commonly_defective_genes = (MUC16, ZNF717, AHNAK2, HYDIN,
HLA-B, PDE4DIP, PKD1L2, MUC3A, OBSCN, PRIM2, HRNR, GPRIN2,
LOC643677, ALPK2, DNHD1, MUC4, PARP4, ZAN, GPR98, MUC6, OR9G9,
AKAP13, AL832834, BDP1, OR4C3, PRUNE2, TTN, COL29A1, DCHS2,
DNAH14, FLJ43860, HLA-DPB1);

next if (grep {$_ eq $gene} @commonly_defective_genes);
```

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Ranking: Functional Impact of Protein-coding Variants

Nonsynonymous SNPs

-exome-wide, pre-calculated datasets for *SIFT, PolyPhen-2, MutationTaster, MutationAssesor, LRT, FATHMM, MetaSVM, MetaLR, SNAP*

-Gain of function: *BSIFT*

Small indels: *InDel SIFT*

Conservation: *PhyloP, GERP++, SiPhy*

All scores normalized to [0-1] (dbNSFP, Liu et al. Human Mutation 2013)

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Ranking: Prior Genes

Generate gene sets (and scores) for genes that are likely to be related to the trait of interest:

- literature search, disease databases, phenoype databases e.g. Epilepsiome
- text mining:
  PubCrawl (normalized Google distance)
  http://pubcrawl.systemsbiology.net
- pathway/network analysis (Ingenuity):
  - topology-related (n-hop neighborhood)
  - enrichment (*EnrichNet*)

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Ranking: GeneAtlas

In-house database for tissue and cell type specific expression data:

- compendium of transcriptome data from ~150 tissue and cell types
- used for prioritize/falsify tissue-specific genes
- integrates multiple public data sources:
  - BodyMap 1.0/2.0 (Illumina RNAseq)
  - various tissue-specific microarray studies (GEO)

The Corpasome

CORPAS family trio WES data

By

Manuel Corpas

http://figshare.com/articles/6_files_with_1GB_per_file/106340

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

CORPAS family WES trio data

- Released under the a CC-BY license, just for issues of compatibility of license.
- “At this point you have permission to use these data in any way you wish as long as you attribute it to the Corpas family.” (from **Manuel Corpas' Blog**)
- For a more detailed explanation please: http://manuelcorpas.com/crowdsourcing/

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

CORPAS family WES trio data

- Fastq files for whole exome sequencing from the Corpas family: mother, father, daughter.
- The data comes from 3 human saliva samples.
- Exome capture was performed using Agilent SureSelect Human All Exon 44
- Sequenced using Illumina's HiSeq technology.

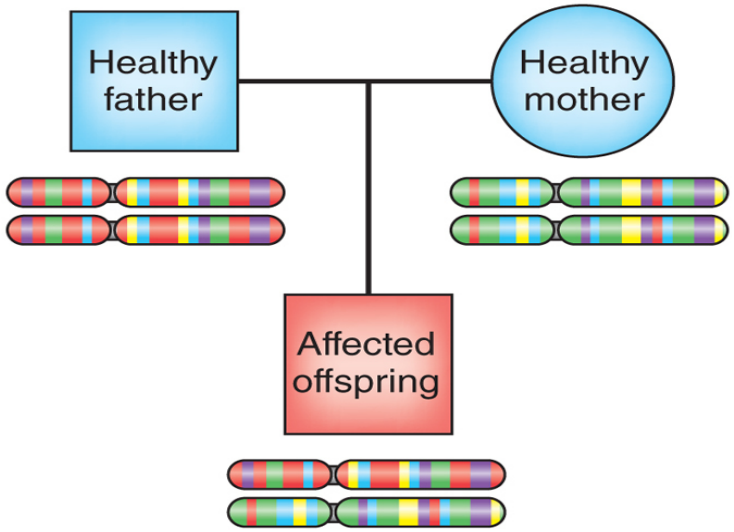
LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

CORPAS family WES trio data

- Only chr22 data
- We introduced two variants related to schizophrenia:
  - one *de novo* mutation only in the child, not in parents
  - one recessive mutation inherited from both parents to the child
- Pedigree/Mode of inheritance (MOI):

	Genotype		
MOI	Father	Mother	Daughter
recessive	01	01	11
de novo	00	00	01



From: Renton & Traynor. Nature Neuroscience, 2013

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

HPC setup:

- Use SSH secure shell
- Login to: student<no>@access-gaia.uni.lu
  - use Port 8022
  - From linux/mac:
    - *ssh –X –p 8022 student<no>@access-gaia.uni.lu*
    - *ssh –X –p 8022 –I $HOME/.ssh/student<no>.key student<no>@access-gaia.uni.lu*
    - *Oarsub –lcore=4,walltime=12 –I*
    - *Copy data from gaia to your local computer:*
      - *cd .ssh*
      - *scp –P 8022 –r –i student<no>.key student<no>@access-gaia.uni.lu:<path-to-file/dir> <destination>*
      - *scp –P 8022 –r –i student<no>.key student<no>@access-gaia.uni.lu:~/WEScourse.tar.gz .*
    - *export LC_ALL=en_US.UTF-8*
- Now you are on the access node of the GAIA cluster of the University Luxembourg (https://hpc.uni.lu)

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

HPC setup:

- NEVER start a program on the access node
- CONNECT to compute nodes (from the access node)
  - START an interactive session:

    oarsub –lcore=4,walltime=12 –I
  - RUN a script

    oarsub <script>

    oarsub –lcore=1,walltime=1 –n sleep ./runscript.sh

- MONITOR your jobs (only works on the access node)

  oarstat –a –u <username>
  - Live status: https://hpc.uni.lu/gaia/monika
- DELETE a job

  oardel <jobid>

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

HPC setup

- Check directory: *pwd*
  - you should be in your home directory
- Untar files:

  tar –zxvf unixtut.tar.gz

  tar –zxvf WEScourse.tar.gz
- Go to course directory:

  cd $HOME/WEScourse

DEMO

Corpasome

Your turn !!

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Demo NGSTools

- FastQC
- BWA
- SAMTOOLS
- PICARD
- GATK
- BEDTOOLS
- ANNOVAR
- DenovoGear
- VCFLIB

+ custom Perl and bash scripts

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Important NGS formats

- fasta
- fastq
- SAM/BAM
- VCF
- BED

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

Other links

- transition/transversion ratio

IGV - Integrative Genomic Viewer

- Download from server
  - *wget http://ipar.ccg.uni-koeln.de/workshop/IGV_2.3.32.zip*
  - *unzip IGV_2.3.32.zip*
  - *cd IGV_2.3.32*
  - *./igv.sh*

- load hg19 genome instead of hg18
- open bam file (first be sure that bam file has an index
  - Run inbefore
    - samtools index <bamfils>

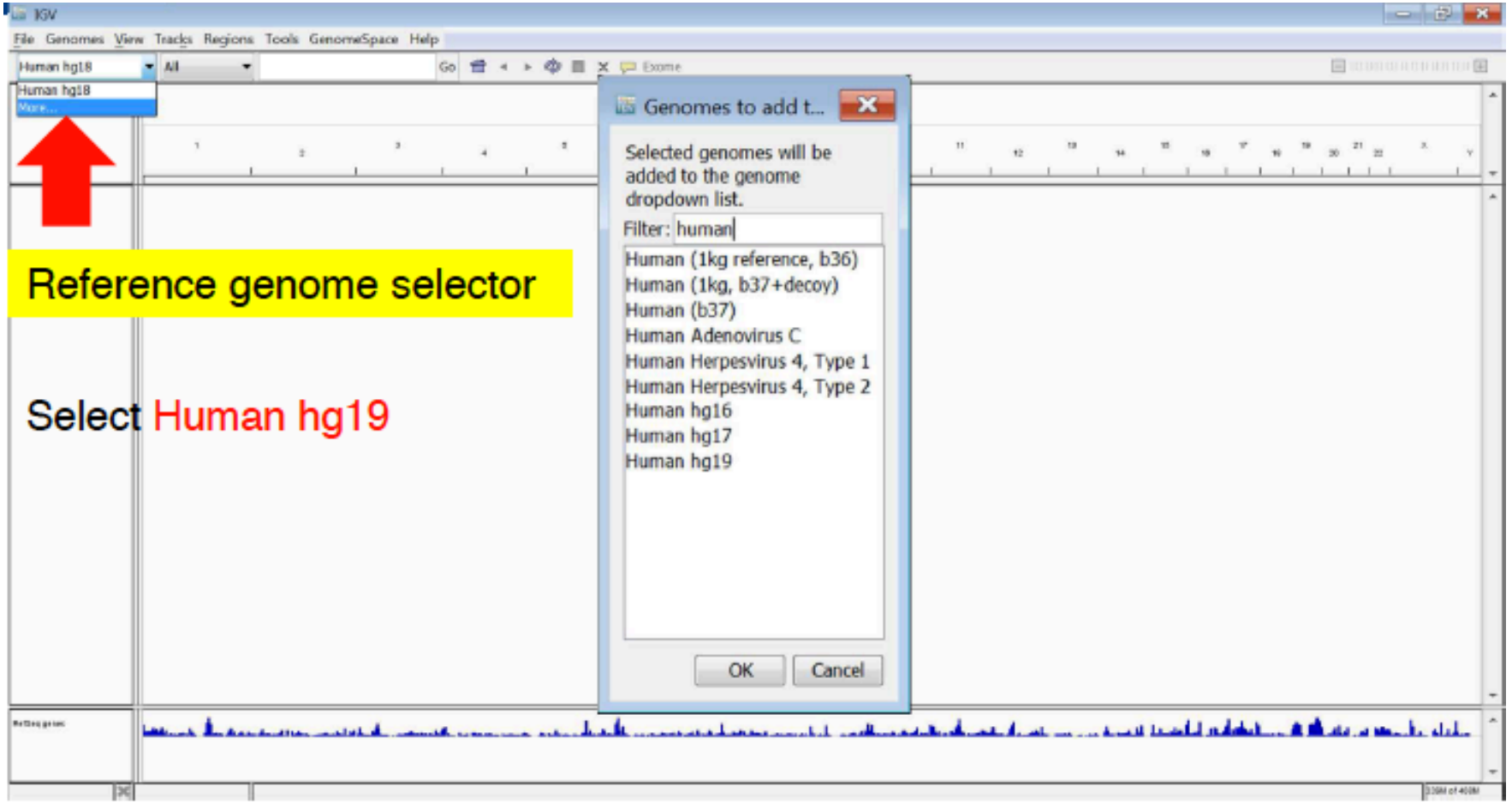
IGV - Integrative Genomic Viewer

- Visualize large genomic data sets on standard
- desktop computers
- NGS read alignments
- Coverage
- Variant
- Splicing of RNA transcripts
- Methylation from bisulfite sequencing
- Microarray data
- Broad institute: IGV is freely available at
  - http://www.broadinstitute.org/igv.

LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

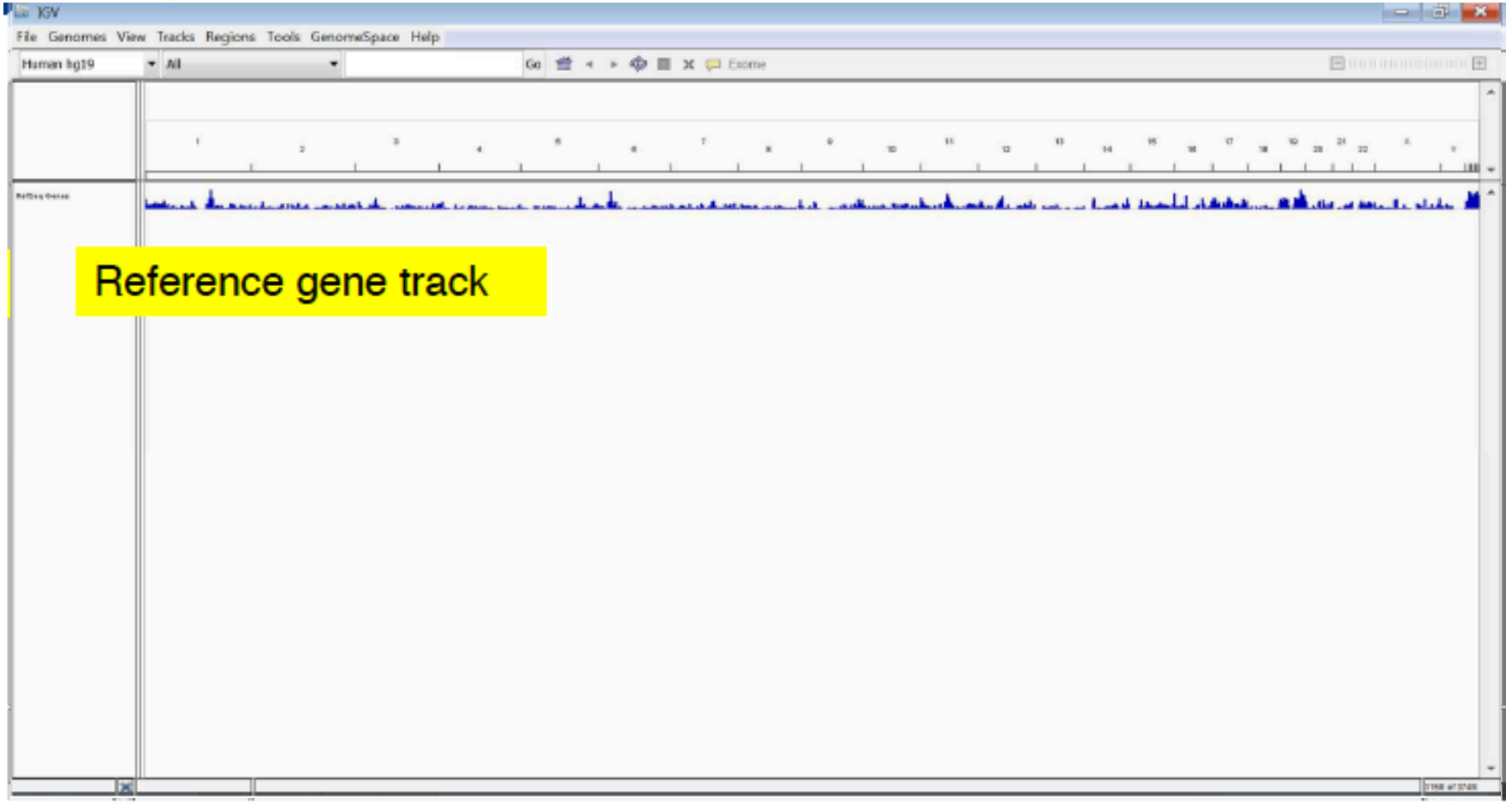
IGV - Select genome (hg19)



LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

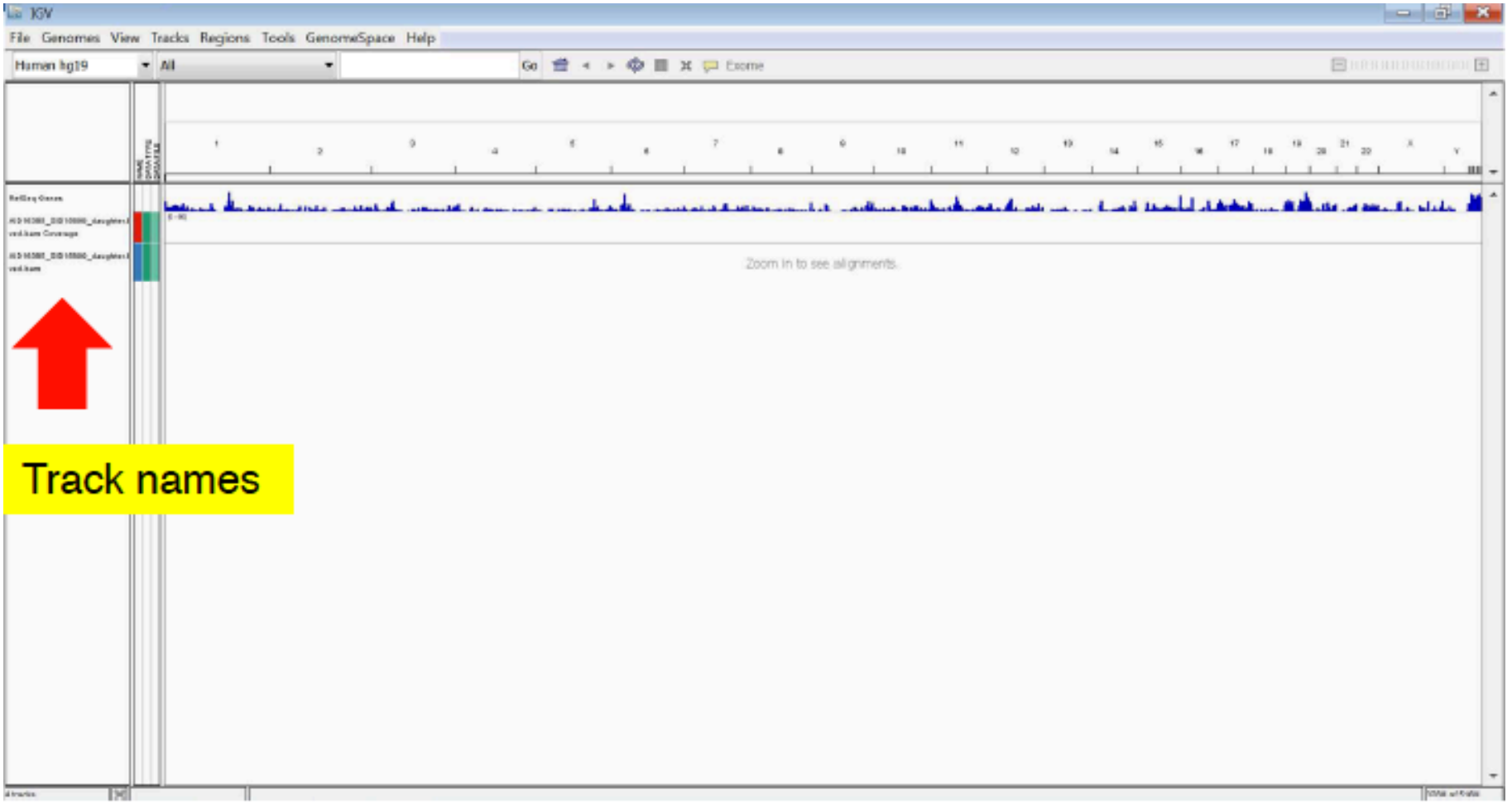
IGV - Select genome (hg19)



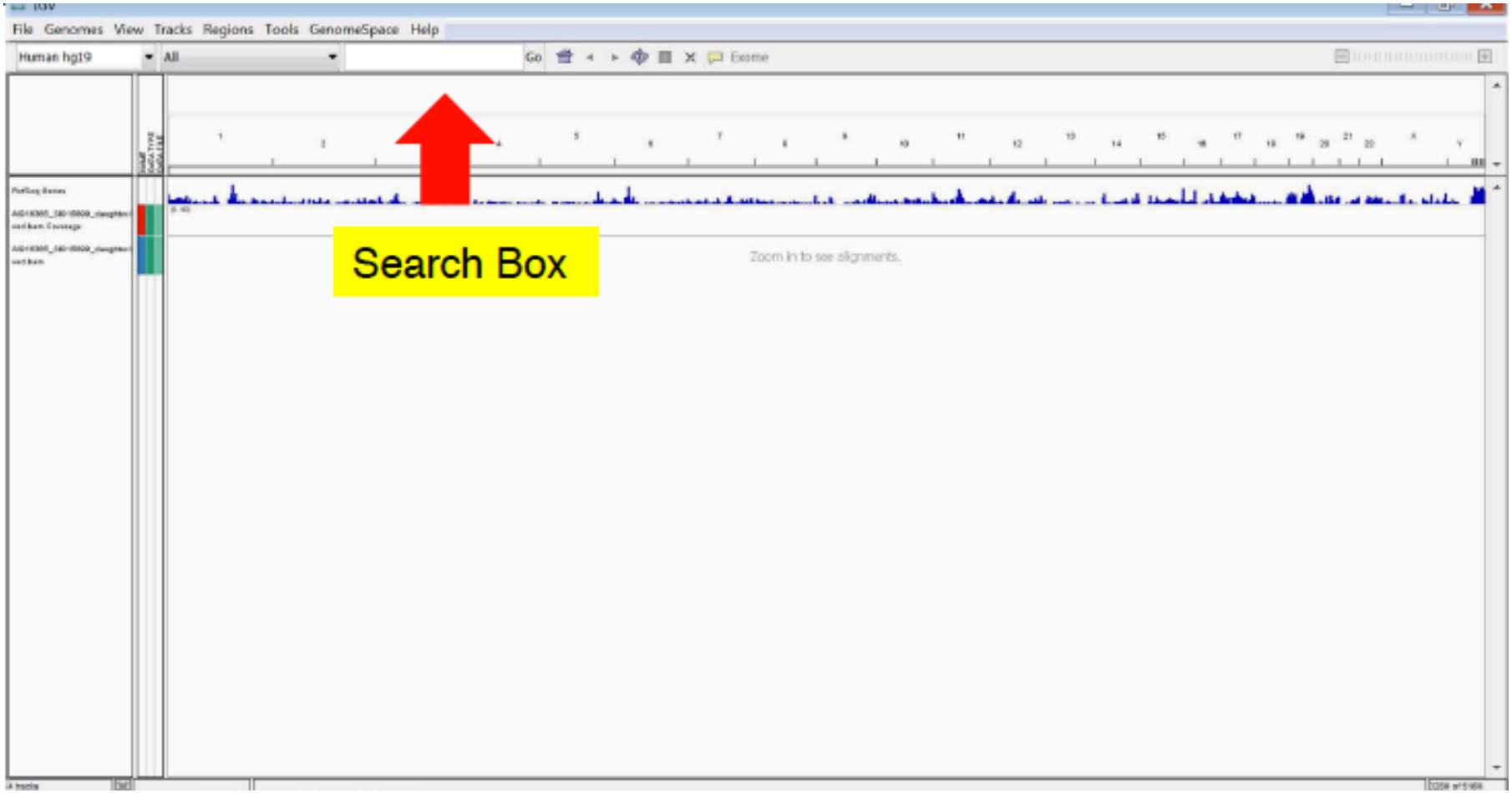
LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

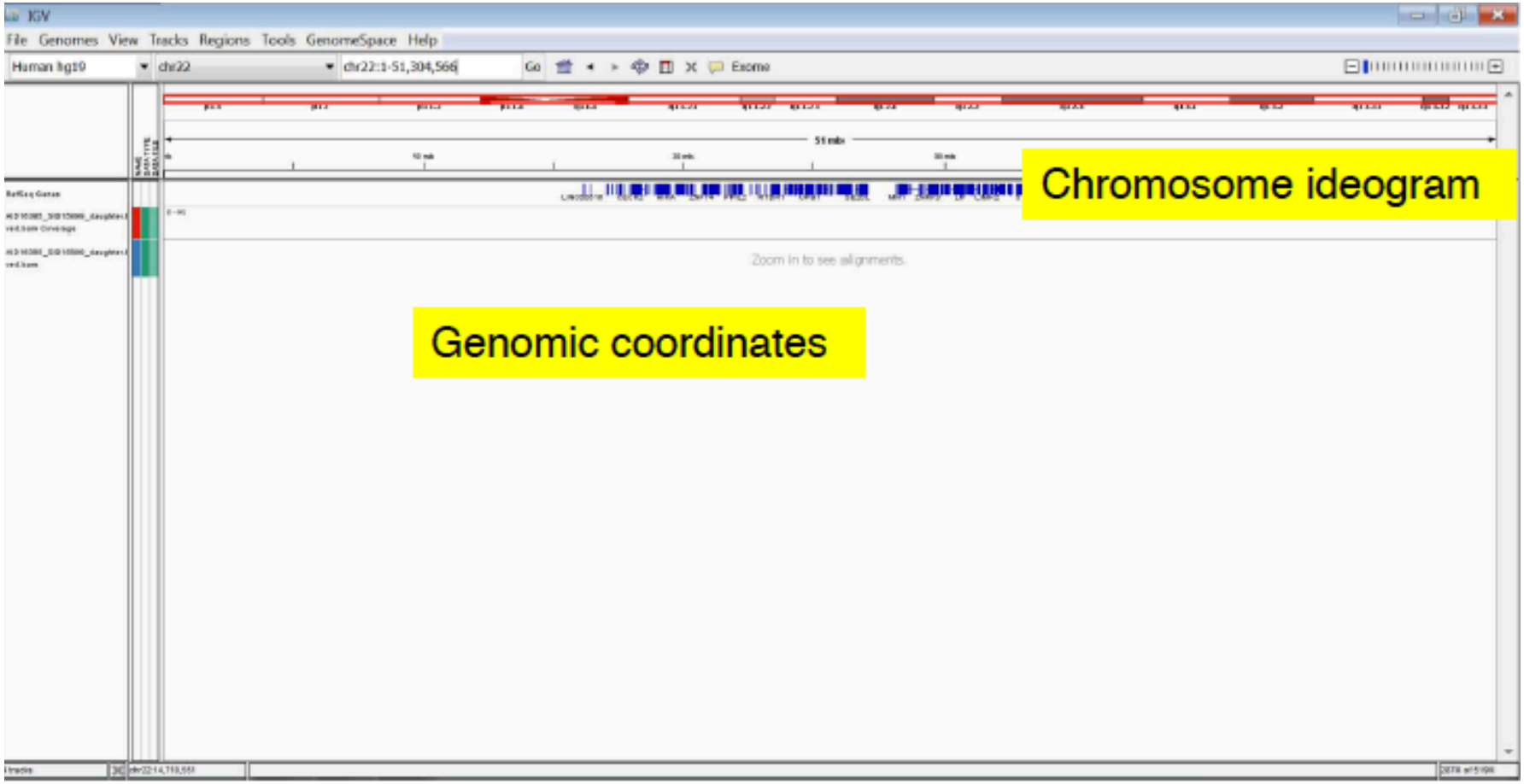
IGV - Tracks


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
All
Go
Zoom in to see alignments.
Track names

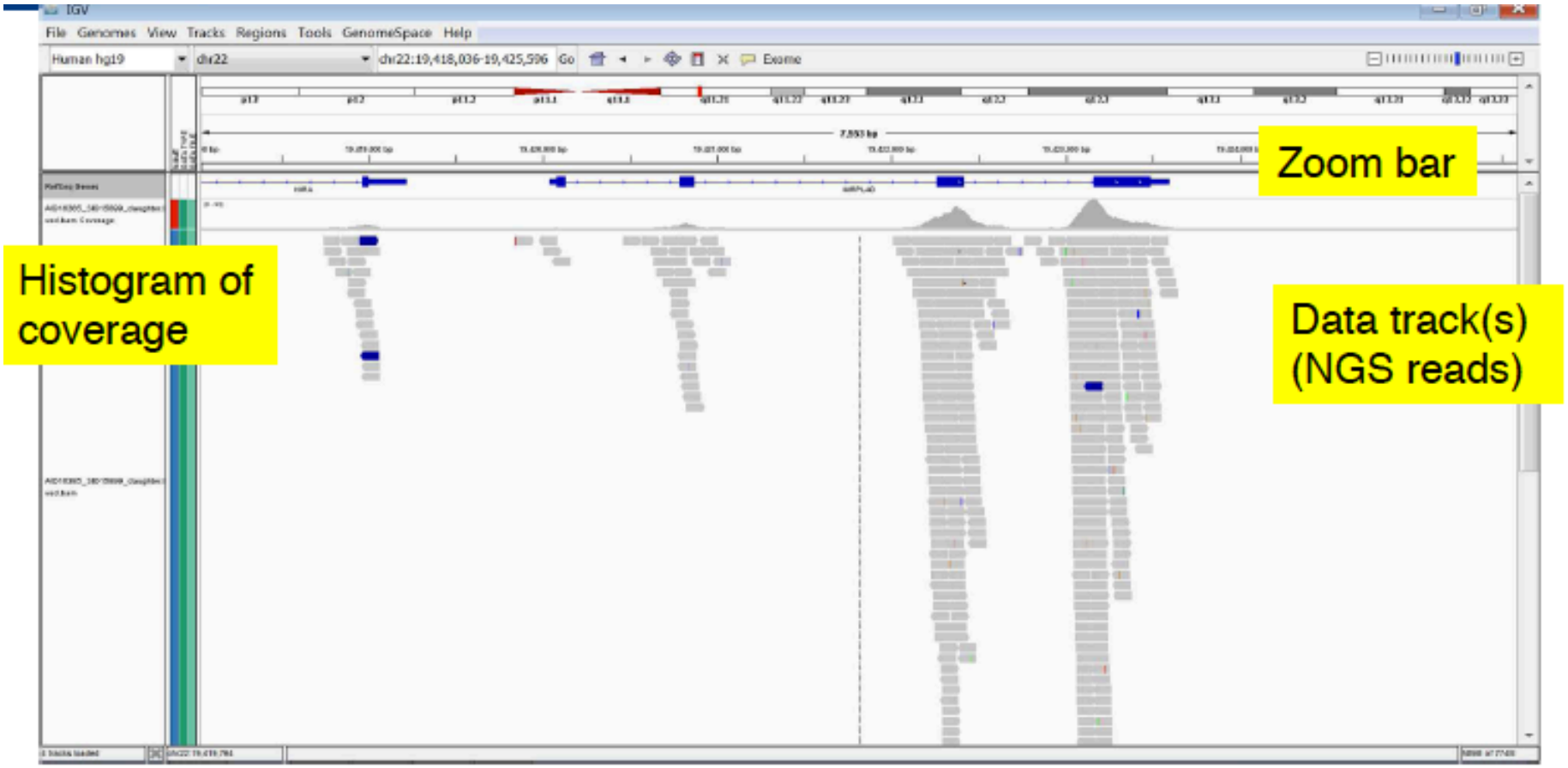
IGV - Search box



IGV - Search box



IGV - Gene of interest MRLP40

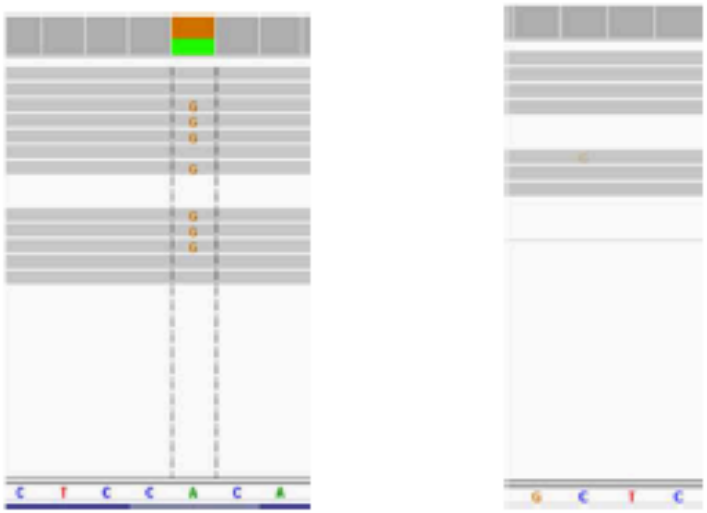


LCSB UNIVERSITÉ DU LUXEMBOURG

Institute for Systems Biology
Revolutionizing Science. Enhancing Life.

IGV - Variants SNPs

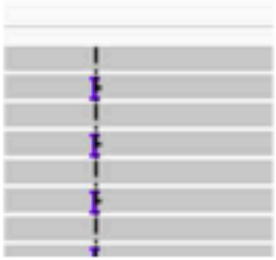
- Gray perfect match with the reference
- Colors – variants



- Dark high quality vs. light low quality

IGV - Variants Indels

- Small insertions



- Small deletions

- For paired end:
  - Red insert size larger than expected
  - Blue insert size smaller than expected

Visualize alignments with samtools tview

```
[hthiele@r715 workshop]$
[hthiele@r715 workshop]$ samtools-0.1.19/samtools tview

Usage: bamtk tview [options] <aln.bam> [ref.fasta]
Options:
   -d display      output as (H)tml or (C)urses or (T)ext
   -p chr:pos      go directly to this position
   -s STR          display only reads from this sample or group


[hthiele@r715 workshop]$
```

- Very simple alignment viewer
- Works on every linux console!
- Quick look up of variants on certain sites

Visualize alignments with samtools tview