

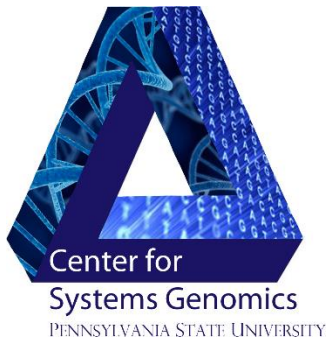
# Predicting genes from DNA Sequence

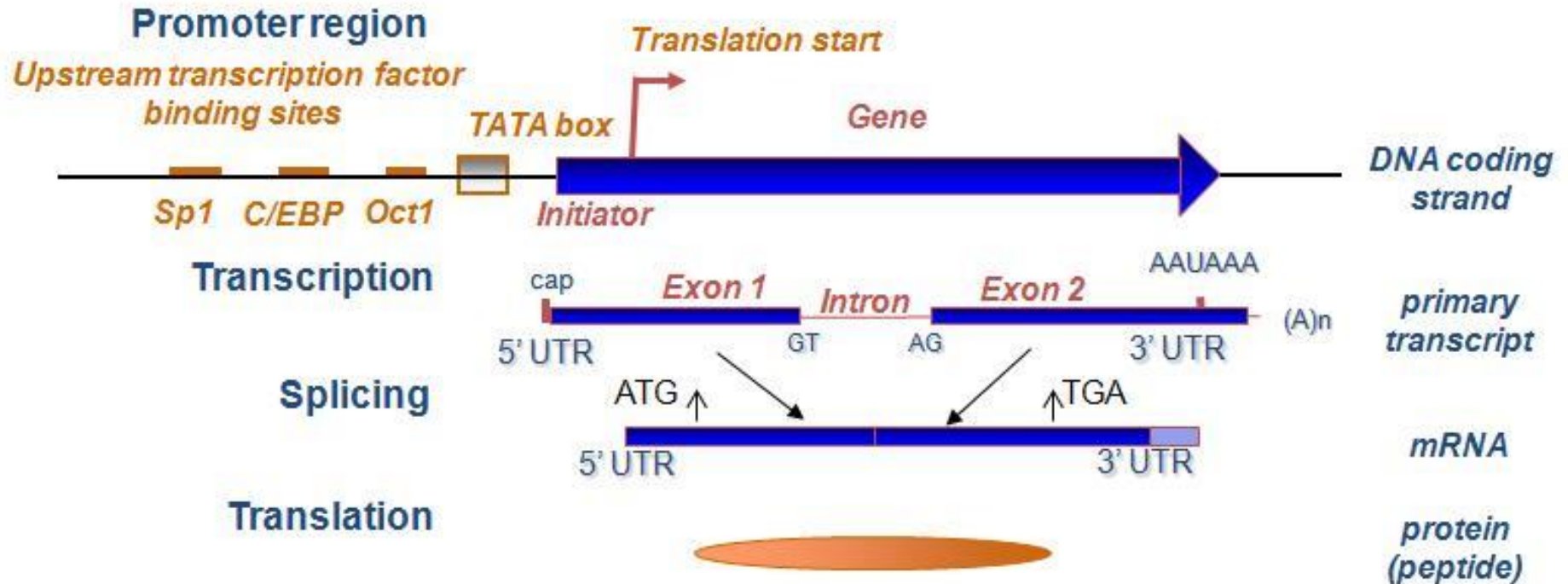
Marylyn D Ritchie, PhD

Professor, Biochemistry and Molecular Biology

Director, Center for Systems Genomics

The Pennsylvania State University





processing this data. An important aspect of complete genomes is the distinction between coding regions and non-coding regions - 'junk' repetitive sequences making up the bulk of base sequences especially in eukaryotes. Within the coding regions, genes are annotated with their translated protein sequence, and often with their cellular function. 2001

According to ENCODE's analysis, 80 percent of the genome has a "biochemical function". More on exactly **what this means later**, but the key point is: **It's not "junk"**. **Scientists have long recognised** that some non-coding DNA has a function, and more and more **solid examples** have **come to light** [*edited for clarity - Ed*]. But, many maintained that much of these sequences were, indeed, junk. ENCODE says otherwise. "Almost every nucleotide is associated with a function of some sort or another, and we now know where they are, what binds to them, what their associations are, and more," says **Tom Gingeras**, one of the study's many senior scientists.

# Gene prediction

- Also called gene finding
- Process of identifying regions of genomic DNA that encode genes
- Includes protein-coding genes and RNA genes
- May also include other functional elements
  - This will be discussed more in the next lecture
  - This area is changing very rapidly



## Gene Prediction

This section includes links to gene prediction programs for both eukaryotic and prokaryotic organisms. Resources that evaluate the available gene predicting programs are also included.

 [RSS Feed](#)  [Compact View](#)  [Sort by Links Directory Index](#)

[DOWNLOAD](#)  [List as XML](#)  [List as JSON](#)  [List as TSV](#)  [List as CSV](#)

 [Hide Resources \(1\)](#)

 [Hide Databases \(2\)](#)

 [Hide Tools \(40\)](#)


**Found 43 links**

**Displaying 15 links**

**AGenDA** 

<http://bibiserv.techfak.uni-bielefeld.de/agenda/> [\[OPEN IN A NEW WINDOW\]](#)

DNA > Gene Prediction

 [Share This Link](#)

Change Text Size: [A](#) [A](#) [A](#)

**Username: \***

**Password: \***

### CAPTCHA

This question is for testing whether you are a human visitor and to prevent automated spam submissions.

**Math question: \***

4 + 0 =

Solve this simple math problem and enter the result. E.g. for 1+3, enter 4.

[Log in](#)

# Empirical gene finding systems

- Uses similarity or homology to identify genes
- Target genome is searched for evidence of similar sequence elements
- Local alignment algorithms look for similarity
  - BLAST
  - FASTA
  - Smith-Waterman

# BLAST

- Basic Local Alignment Search Tool
- Many different types of BLAST available
- First published in 1990, *Journal of Molecular Biology*
- Fast algorithm
- Cannot guarantee optimal alignments – like Smith-Waterman

# BLAST

- Input = sequences in FASTA or GenBank format
- Output = HTML, text, XML
  - Hits found
    - Table showing sequence identifiers for the hits with scoring related data
- BLAST is available for free through NCBI
- Commercial programs that include BLAST are also available



# FASTA

- Originally designed for protein sequence similarity searching
- Added DNA:DNA searches
- Takes a given nucleotide sequence and searches a sequence database to find matches or similar database sequences
- Does a fast search first, followed by Smith-Waterman optimized search

# FASTA

**Label**      **Title Line**      **Comment**

```
>fig|282458.1.peg.1 Chromosomal replication initiator protein dnaA
MSEKEIWEKVL EIAQEKLSAVSYSTFLKDTELYTIKDGEAIVLSSIPFNANWLNQQYAEI
IQAILFDVVG YEVKPHFITTEELANYSNNETATPKEATKPSTETTEDNHVLGREQFNAHN
TFDTEFVIGP GNRFPHAASLAVAEAPAKAYNPLFIYGGVGLGKTHLMHAIGHHVL DNNPDA
KVIYTSSEKFT NEFIKSIRDNEGEAFRERYRNIDVLLIDDIQFIQNKVQTQEEFFYTFNE
LHQNNKQIVISS DRPPKEIAQLEDRLRSRFEWGLIVDITPPDYETRMAILQKKIEEEKLD
IPPEALNYIANQ IQSNIRELEGALTRLLAYSQLLGKPITTELTAEALKDIIQAPKSKKIT
IQDIQKIVGQYY NVRIEDFSAKKRTKSIAYPRQIAMYLSRELTD FSLPKIGE EFGGRDHT
TVIHAHEKISKDL KEDPIFKQEVENLEKEIRNV
```

**Data Lines**

nucleotide sequence  
or  
amino acid sequence

# Ab initio methods

- Intrinsic method based on gene content and signal detection
- DNA sequence is searched for signals of protein-coding genes
  - Promoter sequences
  - One contiguous open reading frame
  - Stop codon
- Complex in higher organisms due to various complexities
- Typically use probabilistic models such as Hidden Markov Models (HMM)
  - GLIMMER, GENSCAN, GenMark

# Combined approaches

- Some combine extrinsic and ab initio approaches
  - Maker
  - Augustus



Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikimedia Shop

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
What links here  
Related changes  
Upload file  
Special pages  
Permanent link  
Page information  
Wikidata item  
Cite this page

Create account Log in

Article Talk

Read Edit View history

Search

## List of gene prediction software

From Wikipedia, the free encyclopedia

*This list is [incomplete](#); you can help by [expanding it](#).*

This **list of gene prediction software** is a compilation of software tools and web portals used for [gene prediction](#).

### *Ab initio* approaches [\[edit\]](#)

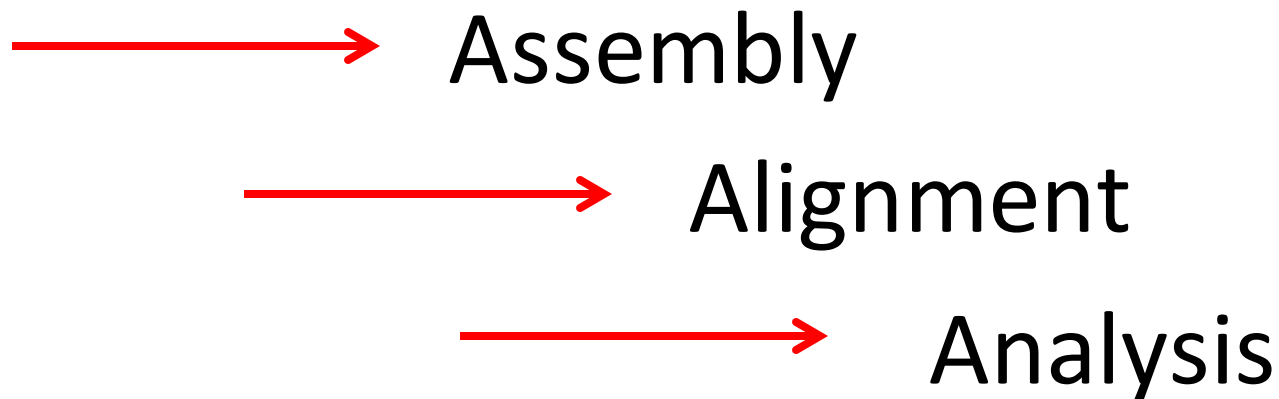
Name	Description	Species	Links	References
<b>ATGpr</b>	identifying translational initiation sites in cDNA sequences			
<b>AUGUSTUS</b>	Eukaryote gene predictor	Eukaryotes	<a href="#">Predict</a> <a href="#">Train</a> <a href="#">AUGUSTUS</a>	<sup>[1]</sup>
<b>BGF</b>	<a href="#">hidden Markov model (HMM)</a> and <a href="#">dynamic programming</a> based <i>ab initio</i> gene prediction program		<a href="#">webserver</a>	
<b>DIODES</b>	a system for fast detection of coding regions in short genomic sequences			
<b>Dragon Promoter Finder</b>	software for recognition of vertebrate RNA Polymerase II promoters			
<b>EUGENE</b>	gene finding for <i>Arabidopsis thaliana</i>	<i>Arabidopsis thaliana</i>		
<b>FGENESH</b>	HMM-based gene structure prediction (multiple genes, both chains)	Eukaryotes	<a href="#">webserver</a>	
<b>FRAMED</b>	find genes and frameshift in G+C rich prokaryotic sequences	Prokaryotes	<a href="#">webserver</a>	<sup>[2]</sup>

When you get your sequence data, what do you do with it?

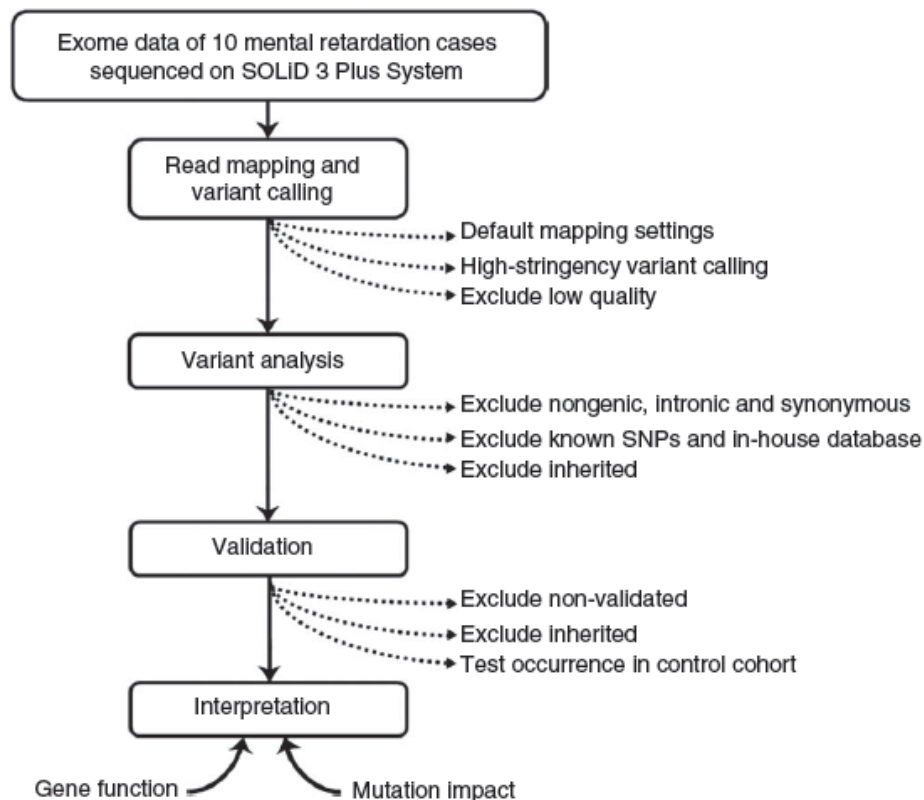
Table 1 | Comparison of next-generation sequencing platforms

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US\$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo- polymer repeats	Bacterial and insect genome <i>de novo</i> assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA <sub>II</sub>	Frag, MP/ solid-phase	RTs	75 or 100	4 <sup>†</sup> , 9 <sup>§</sup>	18 <sup>†</sup> , 35 <sup>§</sup>	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7 <sup>†</sup> , 14 <sup>§</sup>	30 <sup>†</sup> , 50 <sup>§</sup>	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non- cleavable probe SBL	26	5 <sup>§</sup>	12 <sup>§</sup>	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8 <sup>†</sup>	37 <sup>†</sup>	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

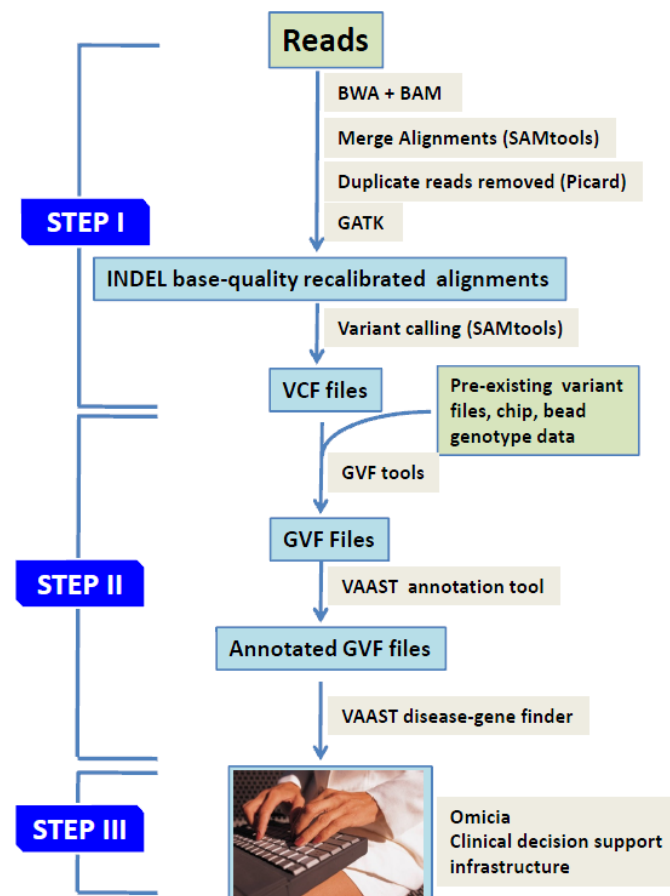
# Sequence data



# Sequence Data Analysis



**Figure 1** Experimental work flow for detecting and prioritizing sequence variants. For all ten mental retardation trios, prioritization of variants observed in the probands was based on selection for non-synonymous changes of high quality only and exclusion of all variants previously observed in healthy individuals, together with those variants that were inherited from an unaffected parent. Interpretation of *de novo* variants was based on gene function and the impact of the mutation.





# Aligning Sequence & Detecting a Sequence Variant

```
AACCGTTAAGACCAAGTCTTTCCGACTCTCGA x 4
ACCGTTAAGACCAAGTCTTTCCGACTCTCGAC x 2
ACCGTTAAGACCAAGTCTTTCCGACTCTCGGC x 2
CCGTTAAGACCAAGTCTTTCCGACTCTCGACT x 1
CGTTAAGACCAAGTCTTTCCGACTCTCGGCTC x 2
GTTAAGACCAAGTCTTTCCGACTCTCGACTCG x 1
GTTAAGACCAAGTCTTTCCGACTCTCGACTCG x 1
TTAAGACCAAGTCTTTCCGACTCTCGACTCGA x 2
TTAAGACCAAGTCTTTCCGACTCTCGGCTCGA x 1
TAAGACCAAGTCTTTCCGACTCTCGACTCGAA x 2
TAAGACCAAGTCTTTCCGACTCTCGGCTCGAA x 2
TAAGACCAAGTCTTTCCGACTCTCGACTCGAA x 1
TAAGACCAAGTCTTTCCGACTCTAGACTCGAA x 1
GACCAAGTCTTTCCGACTCTCGGCTCGAACCT x 1
GACCAAGTCTTTCCGACTCTCGACTCGAACCT x 1
ACCAAGTCTTTCCGACTCTCGACTCGAACCTT x 1
CCAAGTCTTTCCGACTCTCGACTCGAACCTTT x 1
TAAGTCTTTCCGACTCTCTCGGCTCGAACCTTTA x 1
CAAGTCTTTCCGACTCTCGGCTCGAACCTTTA x 1
AAGTCTTTCCGACTCTCGGCTCGAACCTTTAG x 1
AAGTCTTTCCGACTCTCGACTCGAACCTTTAG x 1
AGTCTTTCCGACTCTCGGCTCGAACCTTTAGG x 1
GTCTTTCCGACTCTCGACTCGAACCTTTAGGT x 1
GTCTTTCCGACTCTCGGCTCGAACCTTTAGGT x 1
TCTTTCCGACTCTCGGCTCGAACCTTTAGGTG x 2
TCTTTCCGACTCTCGACTCGAACCTTTAGGTG x 1
CTTTCCGACTCTCGACTCGAACCTTTAGGTGT x 1
CTTTCCGACTCTCTCGGCTCGAACCTTTAGGTGT x 1
TTTCCGACTCTCGACTCGAACCTTTAGGTGTA x 2
TTTCCGACTCTCGGCTCGAACCTTTAGGTGTA x 1
TTCCGACTCTCGACTCGAACCTTTAGGTGTAA x 2
TCCGACTCTCGACTCGAACCTTTAGGTGTAAA x 3
CGGACTCTCGGCTCGAACCTTTAGGTGTAAAA x 1
CGGACTCTCGGCTCGAACCTTTAGGTGTAAAA x 1
GGAACCTCTCGGCTCGAACCTTTAGGTGTAAAAAG x 1
GACTCTCGGCTCGAACCTTTAGGTGTAAAAAGA x 1
ACTCTCGACTCGAACCTTTAGGTGTAAAAAGAG x 1
CTCTCGGCTCGAACCTTTAGGTGTAAAAAGAGA x 1
CTCTCGACTCGAACCTTTAGGTGTAAAAAGAGA x 1
CTCGACTCGAACCTTTAGGTGTAAAAAGAGACC x 1
TCGACTCGAACCTTTAGGTGTAAAAAGAGACCG x 2
TCGGCTCGAACCTTTAGGTGTAAAAAGAGACCG x 1
CGGCTCGAACCTTTAGGTGTAAAAAGAGACCGA x 1
TTGGCAATTCTGGTTTCAGAAAGCCTGAGAGCCGAGCTTGGAAATCCACATTTTCTCTGGCTGC
```

**Figure 1** Detection of sequence variants. A total of 32 nucleotide NGS reads (top, sequence mismatches in red) aligned with the genomic reference sequence (bottom). The center of the alignment shows a variant present in the heterozygous state. '×*n*' behind the read indicates how many identical reads were obtained.

## Human Chromosomes 2 and 20: 1.591 kbp from chr20:67,960..69,550

### Instructions

Search using a sequence name, gene name, locus, or other landmark. The wildcard character \* is allowed.

Navigate by clicking one of the rulers to center on a location, or click and drag to select a region. Use the Scroll/Zoom buttons to change magnification and position.

Examples: chr20:67960..69559, chr2:2043960..2045540.

[\[Bookmark this\]](#) [\[Add custom tracks\]](#) [\[Share these tracks\]](#) [\[Send to Galaxy\]](#) [\[Link to Image\]](#) [\[High-res Image\]](#) [\[Help\]](#) [\[Reset\]](#) [\[Make an Error\]](#)

### Search

Landmark or Region:

chr20:67960..69550

Search

Annotate Restriction Sites

Configure...

Go

Data Source

Human Chromosomes 2 and 20

Scroll/Zoom:



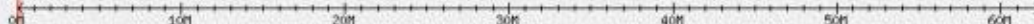
Show 1.591 kbp



☐ Flip

### Overview

chr20



### Region

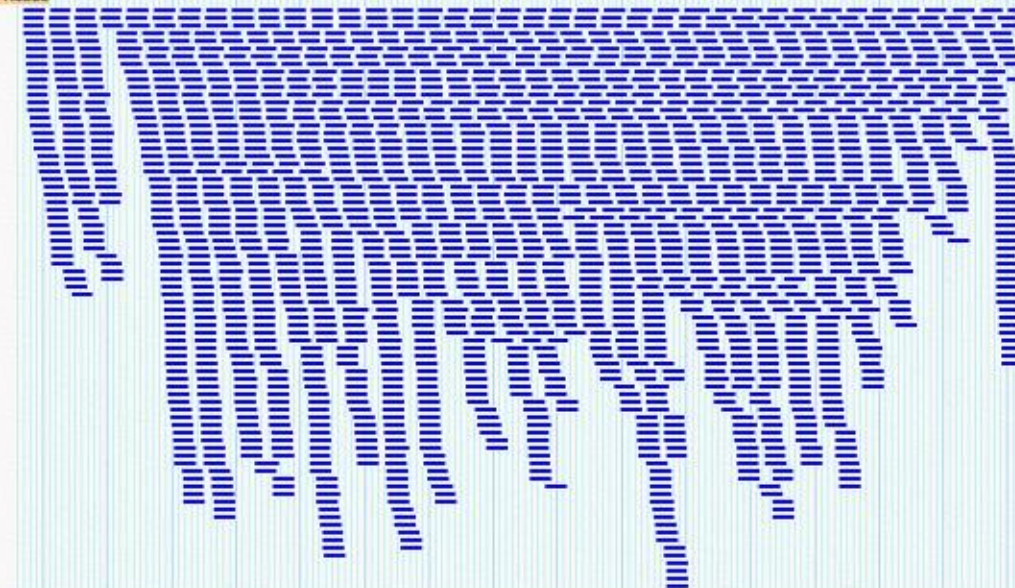


### Details



### Ensembl predicted genes

### Reads



### Tracks

☒ Basic features ☐ All on ☐ All off

☐ 6-frame translation

☐ DNA

☒ Ensembl predicted genes

☒ Reads ☐ All on ☐ All off

☐ Coverage (density plot)

☐ Coverage (xyplot)

☒ Reads

☒ Analysis ☐ All on ☐ All off

☐ plugin:Restriction Sites

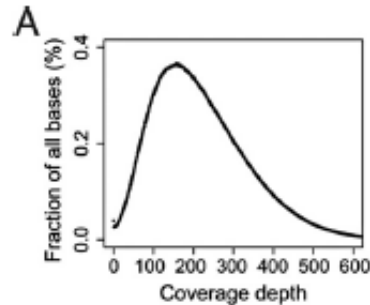
☒ Add custom tracks

[\[Help\]](#)

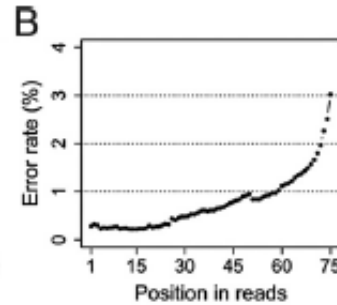
Brave

# Technical Considerations for NGS

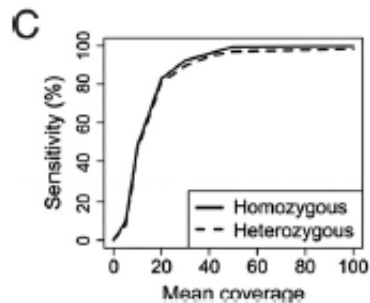
Coverage depth of targeted bases



Per base error rate related to read position



Sensitivity : Coverage depth vs detection of heterozygous variants



Sensitivity of detection of the variants at exact per base coverage

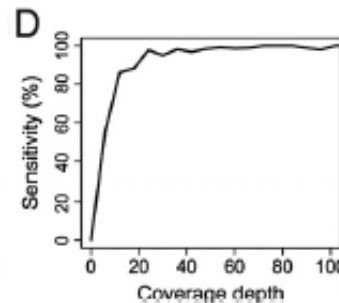
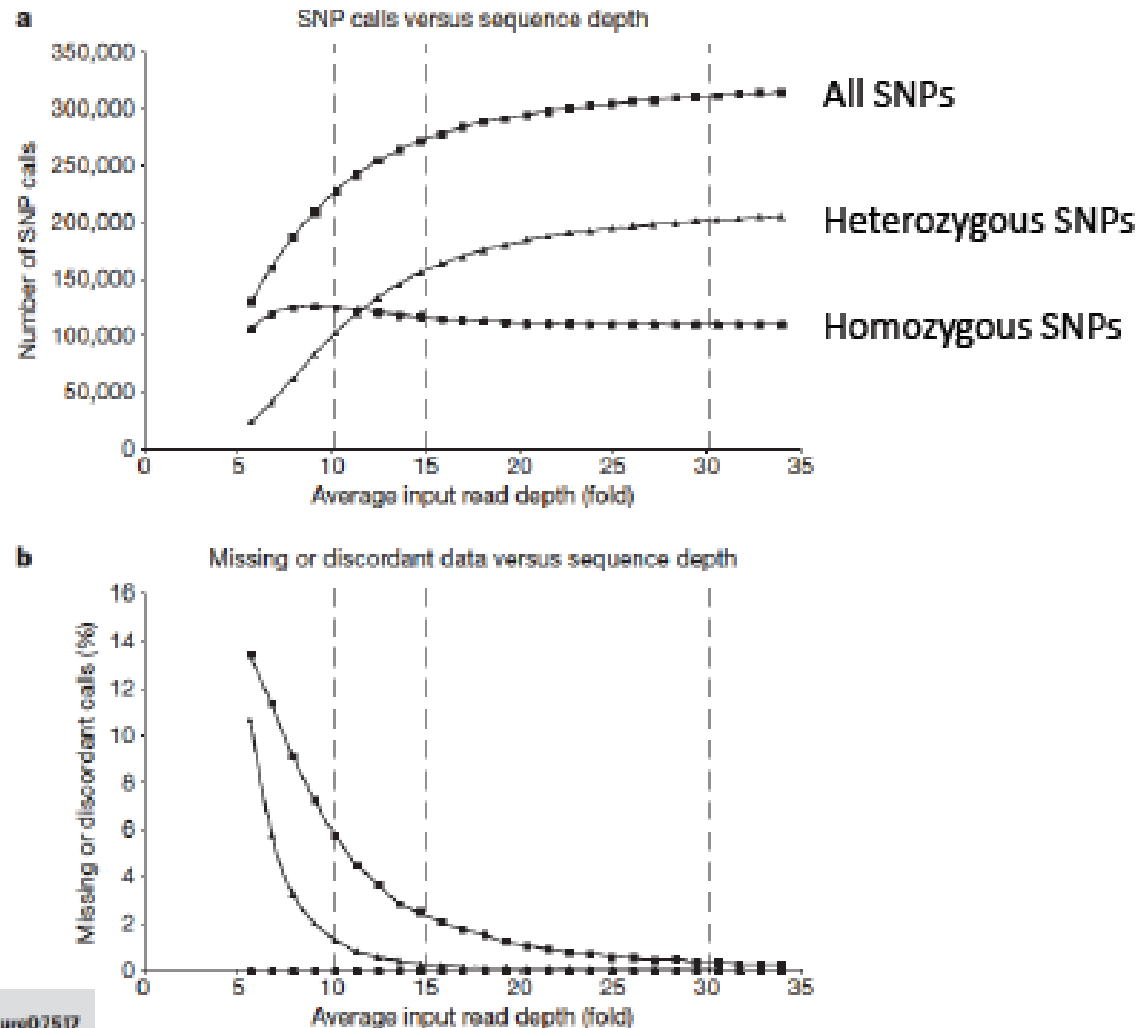


Fig. 1. Coverage of targeted bases, error rate, and sensitivity to detect variants in whole-exome capture data. (A) Distribution of per-base read coverage among 5 capture experiments. A small fraction of targeted bases are poorly captured across all experiments. (B) The per-base error rate in this data set is shown as a function of read position. (C) Subject GIT 264-1 was sequenced to a mean depth of 99x. The sensitivity to detect homozygous (solid line) or heterozygous (dashed line) variants as mean depth of whole-exome sequence coverage increases from 0 to 100x is shown. Sensitivity to detect heterozygous variants increases from 81% to 90% to 95% as mean coverage is increased from 20x to 30x and 40x, and plateaus at 98%. (D) Sensitivity of detection of heterozygous variants at exact per-base coverage. Sensitivity is approximately 80% at 10x coverage, and approaches 100% at or greater than 20x per-base coverage.

- Different technologies have different limitations.

# Why 30X Coverage?

## Why 30X?





# NGS Quality Score

Probability of  
Incorrect Base Call (e)

Base Call Accuracy

Quality Score

1 in 10

90 %

10

1 in 100

99 %

20

1 in 1000

99.9 %

30

1 in 10000

99.99 %

40

NGS

Coverage

\_\_\_\_\_

# Galaxy

- <http://main.g2.bx.psu.edu/>



Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the [free public server](#) or [your own instance](#), you can perform, reproduce, and share complete analyses.

## Use Galaxy



[Use the free public server](#)

## Get Galaxy



Install [locally](#) or [in the cloud](#)

## Learn Galaxy



[Screencasts](#), [Galaxy 101](#),  
[...](#)

## Get Involved



[Mailing lists](#), [Tool Shed](#),  
[wiki](#)

The [Galaxy Team](#) is a part of [BX](#) at [Penn State](#), and the [Biology](#) and [Mathematics and Computer Science](#) departments at [Emory University](#). The Galaxy Project is supported in part by [NSF](#), [NHGRI](#), [The Huck Institutes of the Life Sciences](#), [The Institute for CyberScience](#) at Penn State,

**Table I:** List of main web-based general genome browsers with multiple species

Name	URL	Description
Ensembl	<a href="http://www.ensembl.org/">http://www.ensembl.org/</a>	Major species with completed genome sequences providing lineage-specific web portals for vertebrates, metazoa, plants, fungi, protists and bacteria.
UCSC	<a href="http://genome.ucsc.edu/cgi-bin/hgGateway">http://genome.ucsc.edu/cgi-bin/hgGateway</a>	Major species with completed genome sequences including vertebrates, deuterostomes, insects and nematodes. No plant species.
Map Viewer	<a href="http://www.ncbi.nlm.nih.gov/mapview/">http://www.ncbi.nlm.nih.gov/mapview/</a>	Major species with completed genome sequences including vertebrates, invertebrates, protozoa, plants and fungi, as well as dozens of uncompleted plant genomes.
Phytozome	<a href="http://www.phytozome.net/cgi-bin/gbrowse/">http://www.phytozome.net/cgi-bin/gbrowse/</a>	Major plant species with completed and ongoing genome sequences including monocots, dicots, fern, moss and green algae, with VISTA alignments.
Gramene	<a href="http://www.gramene.org/genome.browser/">http://www.gramene.org/genome.browser/</a>	Major plant species with completed genome sequences including monocots, dicots, fern, moss and green algae, and the short arm of chromosome 3 of several wild rice species.
VISTA	<a href="http://pipeline.lbl.gov/cgi-bin/gateway2/">http://pipeline.lbl.gov/cgi-bin/gateway2/</a>	Whole genome alignment presentation, including vertebrates, insects, nematodes, deuterostomes, plants, fungi, alga, annelids, stramenopiles and metazoa.
Genome Projector	<a href="http://www.g-language.org/g3/">http://www.g-language.org/g3/</a>	Several hundreds of bacteria genomes with circular or linear maps.
Annmap	<a href="http://annmap.picr.man.ac.uk/">http://annmap.picr.man.ac.uk/</a>	A genome browser that includes mappings between genomic features and Affymetrix microarrays for human, mouse, rat and yeast.

**Table 2:** List of some web-based species-specific genome browsers

Name	URL	Species
<b>Animals</b>		
MGI	<a href="http://gbrowse.informatics.jax.org/cgi-bin/gbrowse/">http://gbrowse.informatics.jax.org/cgi-bin/gbrowse/</a>	<i>Mus musculus</i> (Mouse)
RGD	<a href="http://rgd.mcw.edu/fgb2/gbrowse/">http://rgd.mcw.edu/fgb2/gbrowse/</a>	<i>Rattus norvegicus</i> (Rat)
Xenbase	<a href="http://www.xenbase.org/fgb2/gbrowse/">http://www.xenbase.org/fgb2/gbrowse/</a>	<i>Xenopus tropicalis</i> (Frog)
ZFIN	<a href="http://zfin.org/cgi-perl/gbrowse/">http://zfin.org/cgi-perl/gbrowse/</a>	<i>Danio rareo</i> (Zebrafish)
Flybase	<a href="http://flybase.org/cgi-bin/gbrowse/">http://flybase.org/cgi-bin/gbrowse/</a>	<i>Drosophila</i> (Fruit fly)
BeetleBase	<a href="http://beetlebase.org/cgi-bin/gbrowse/">http://beetlebase.org/cgi-bin/gbrowse/</a>	<i>Tribolium Castaneum</i> (Beetle)
AphidBase	<a href="http://isyip.genouest.org/cgi-bin/gb2/gbrowse/">http://isyip.genouest.org/cgi-bin/gb2/gbrowse/</a>	<i>Acyrtosiphon pisum</i> (Aphid)
wFleaBase	<a href="http://wfleabase.org/gbrowse/">http://wfleabase.org/gbrowse/</a>	<i>Daphnia</i> (Water flea)
Wormbase	<a href="http://www.wormbase.org/db/gb2/gbrowse/">http://www.wormbase.org/db/gb2/gbrowse/</a>	<i>Caenorhaditus elegans</i> (Worm)
<b>Plants</b>		
TAIR	<a href="http://www.arabidopsis.org/browse/">http://www.arabidopsis.org/browse/</a>	<i>Arabidopsis thaliana</i> (Wall cress)
BRAD	<a href="http://brassicadb.org/cgi-bin/gbrowse/">http://brassicadb.org/cgi-bin/gbrowse/</a>	<i>Brassica rapa</i> (Brassica)
SGN	<a href="http://solgenomics.net/gbrowse/bin/gbrowse/">http://solgenomics.net/gbrowse/bin/gbrowse/</a>	<i>Solanum pimpinellifolium</i> (Tomato)
Popgenie	<a href="http://www.popgenie.org/tool/gbrowse/">http://www.popgenie.org/tool/gbrowse/</a>	<i>Populus trichocarpa</i> (Populus)
Rice Genome	<a href="http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/">http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/</a>	<i>Oryza sativa japonica</i> (Rice)
Rice-Map	<a href="http://www.ricemap.org/">http://www.ricemap.org/</a>	<i>Oryza sativa japonica/indica</i> (Rice)
MaizeDB	<a href="http://gbrowse.maizegdb.org/">http://gbrowse.maizegdb.org/</a>	<i>Zea mays</i> (Maize)
<b>Microbes</b>		
dictyBase	<a href="http://dictybase.org/db/cgi-bin/ggb/gbrowse/">http://dictybase.org/db/cgi-bin/ggb/gbrowse/</a>	<i>Dictyostelium discoideum</i> (Dictyostelid)
SGD	<a href="http://browse.yeastgenome.org/">http://browse.yeastgenome.org/</a>	<i>Saccharomyces cerevisiae</i> (Yeast)
ParameciumDB	<a href="http://paramecium.cgm.cnrs-gif.fr/cgi-bin/gbrowse2/">http://paramecium.cgm.cnrs-gif.fr/cgi-bin/gbrowse2/</a>	<i>Paramecium tetraurelia</i>




## Genome Information by organism

[Download Reports from FTP s](#)

[Overview \[10523\]](#)
[Eukaryotes \[1678\]](#)
[Prokaryotes \[28079\]](#)
[Viruses \[4195\]](#)
[Plasmids \[5012\]](#)

[Download selected records](#)

Items 1 - 100 of 10523 << First < Prev Page 1 of 106 Next > Last >>								
Organism/Name	Kingdom	Group	SubGroup	Size (Mb)	Chr	Organelles	Plasmids	Assemblies
	All ▼	All ▼	All ▼					
'Chrysanthemum coronarium' phytoplasma	Bacteria	Tenericutes	Mollicutes	0.739592	-	-	-	1
Abaca bunchy top virus	Viruses	ssDNA viruses	Nanoviridae	0.006422	6	-	-	1
Abalone herpesvirus Victoria/AUS/2009	Viruses	dsDNA viruses, no RNA stage	unclassified	0.211518	1	-	-	1
Abalone shriveling syndrome-associated virus	Viruses	dsDNA viruses, no RNA stage	unclassified	0.034952	1	-	-	1
Abelson murine leukemia virus	Viruses	Retro-transcribing viruses	Retroviridae	0.005894	1	-	-	1
Abiotrophia defectiva	Bacteria	Firmicutes	Bacilli	2.04344	-	-	-	1
Abutilon Brazil virus	Viruses	ssDNA viruses	Geminiviridae	0.005271	2	-	-	1
Abutilon mosaic Bolivia virus	Viruses	ssDNA viruses	Geminiviridae	0.005399	2	-	-	1
Abutilon mosaic Brazil virus	Viruses	ssDNA viruses	Geminiviridae	0.005282	2	-	-	1
Abutilon mosaic virus	Viruses	ssDNA viruses	Geminiviridae	0.005217	2	-	-	1
Acanthamoeba castellanii	Eukaryota	Protists	Other Protists	46.7146	-	1	-	2
Acanthamoeba polyphaga mimivirus	Viruses	dsDNA viruses, no RNA stage	Mimiviridae	1.18155	1	-	-	1
Acanthisitta chloris	Eukaryota	Animals	Birds	1035.88	-	-	-	1
Acanthocystis surfacea Chlorella virus 1	Viruses	dsDNA viruses, no RNA stage	Phycodnaviridae	0.288047	1	-	-	1

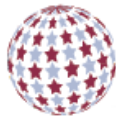

**Galaxy**

[Analyze Data](#)
[Workflow](#)
[Shared Data](#)
[Visualization](#)
[Cloud](#)
[Help](#)
[User](#)

Tools


[Get Data](#)  
[Lift-Over](#)  
[Text Manipulation](#)  
[Convert Formats](#)  
[FASTA manipulation](#)  
[Filter and Sort](#)  
[Join, Subtract and Group](#)  
[Extract Features](#)  
[Fetch Sequences](#)  
[Fetch Alignments](#)  
[Get Genomic Scores](#)  
[Operate on Genomic Intervals](#)  
[Statistics](#)  
[Graph/Display Data](#)  
[Regional Variation](#)  
[Multiple regression](#)  
[Multivariate Analysis](#)  
[Evolution](#)  
[Motif Tools](#)  
[Multiple Alignments](#)  
[Metagenomic analyses](#)  
[Genome Diversity](#)  
  
[NGS TOOLBOX BETA](#)

**Galaxy** is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy [start here](#) or consult our [help resources](#).




Want help?  
Get answers.

**Biostars**  
GALAXY EXPLAINED




Tweets



**Galaxy Project** @galaxyproject 1h


CDD Ingénieur NGS - Institut Curie, Paris [bit.ly/YGIDHy](http://bit.ly/YGIDHy) #usegalaxy



**Galaxy Project** @galaxyproject 12h





Have you ever wanted to improve a #usegalaxy Tool? For a large number of tools, that just got easier: [github.com/galaxyproject/...](https://github.com/galaxyproject/)

Show Summary



**Galaxy Project** @galaxyproject 15h


Everyone's working on data analysis challenge: The 1800th Galaxy CiteULike paper is in Requirements Engineering [bit.ly/gxyCiteULike](http://bit.ly/gxyCiteULike)

<https://usegalaxy.org/>

<https://www.biostars.org>
[Twitter](#)
[Facebook](#)
[Pinterest](#)
[YouTube](#)
[Instagram](#)
[LinkedIn](#)
[Ritchie Lab Blogs](#)
[Ritchie Lab Wiki](#)
[Ritchie Lab Forum](#)
[PASS Explorer](#)
[PSU-Websites](#)

[LATEST](#)
[OPEN](#)
[RNA-SEQ](#)
[CHIP-SEQ](#)
[SNP](#)
[ASSEMBLY](#)
[TUTORIALS](#)
[TOOLS](#)
[JOBS](#)


**Biostars**  
 — BIOINFORMATICS EXPLAINED —

Welcome to Biostar!
 [Community](#)
[User Login](#)
[New Post](#)

Live search: start typing...
 or [Q C](#)

Limit to: all time ▾
 <prev • 18,880 results • page 1 of 472 • next >
 Sort by: update ▾

0 votes	1 answer	42 views	<a href="#">trim leading T or A from fastq file</a> <a href="#">rna-seq</a>	written 35 minutes ago by <a href="#">meishengxiao86</a> • 0 • updated 7 minutes ago by <a href="#">Pierre Lindenbaum</a> ♦ 63k
12 votes	1 answer 6 follow	934 views	<a href="#">Forum: Free Ensembl Courses</a> <a href="#">ensembl</a> <a href="#">forum</a> <a href="#">genome-browser</a>	written 15 months ago by <a href="#">Emily_Ensembl</a> ♦ 4.6k
0 votes	0 answers	30 views	<a href="#">Trouble with samtools mpileup on my exome bam files</a> <a href="#">alignment</a> <a href="#">snp</a> <a href="#">sequencing</a>	written 22 minutes ago by <a href="#">vchris_ngs</a> • 110
0 votes	0 answers	28 views	<a href="#">Circular-Linear Regression In R: Error in while (diff &gt; tol) { : missing value where TRUE/FALSE needed</a> <a href="#">R</a>	written 23 minutes ago by <a href="#">fx2038</a> • 0
0 votes	2 answers	146 views	<a href="#">How to deal with demultiplexed Miseq pair-end (2*250bp) 16S data using QIIME?</a> <a href="#">next-gen</a> <a href="#">sequencing</a>	written 1 day ago by <a href="#">nkuyfq</a> • 0 • updated 2 hours ago by <a href="#">sheridan.christopher</a> • 0

# Questions???

