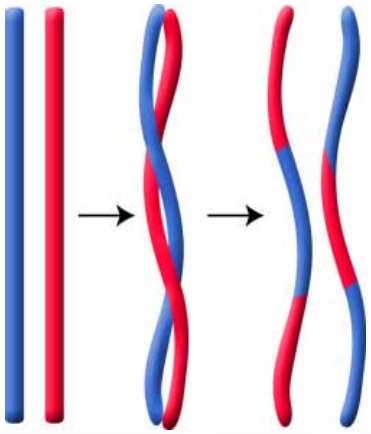


NEXT GENERATION SEQUENCING
Applications & Analysis

Petrus Tang, Ph.D. (鄧致剛)
Graduate Institute of Basic Medical Sciences
and
Bioinformatics Center, Chang Gung University.
petang@mail.cgu.edu.tw
http://petang.cgu.edu.tw

Applications on Biomedical Sciences


ABI SOLiD 5500XL


Illumina Hiseq 2000


Roche 454 FLX


ABI Ion-Torrent

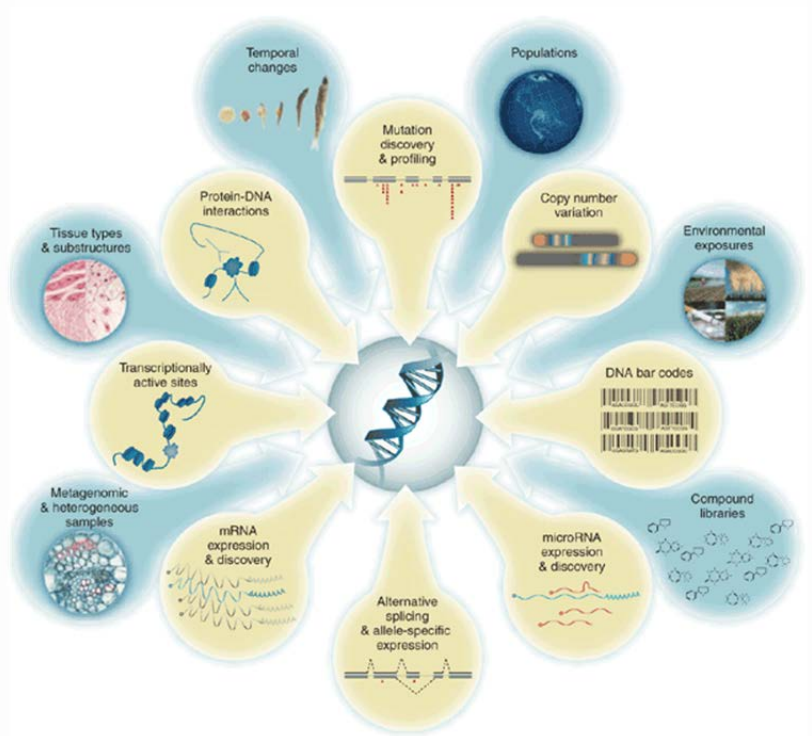

Illumina Miseq


ABI 3730XL

DNA
- Exome Sequencing
- Genome Resequencing
- De novo Genome Sequencing
- Meta-Genomic Sequencing
- Chip-Sequencing

RNA
- Small RNA Sequencing
- Transcriptome Sequencing
- De novo Transcriptome Sequencing
- Meta-Transcriptome Sequencing

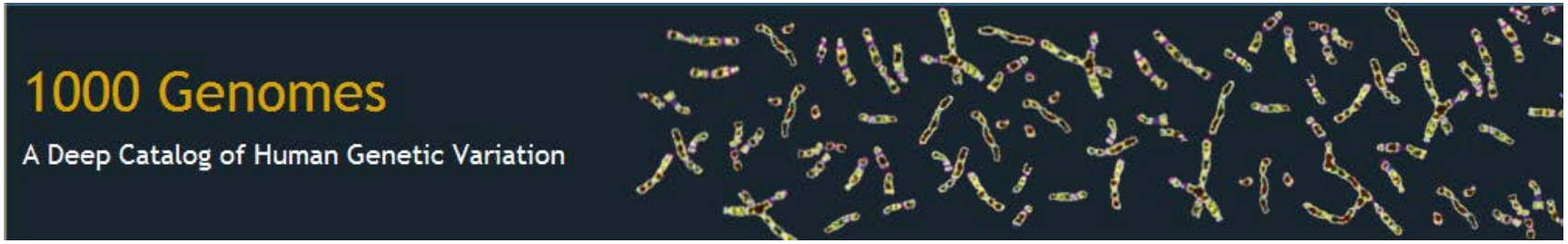
BIOINFORMATICS




IBM X3850

Applications on Biomedical Sciences

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

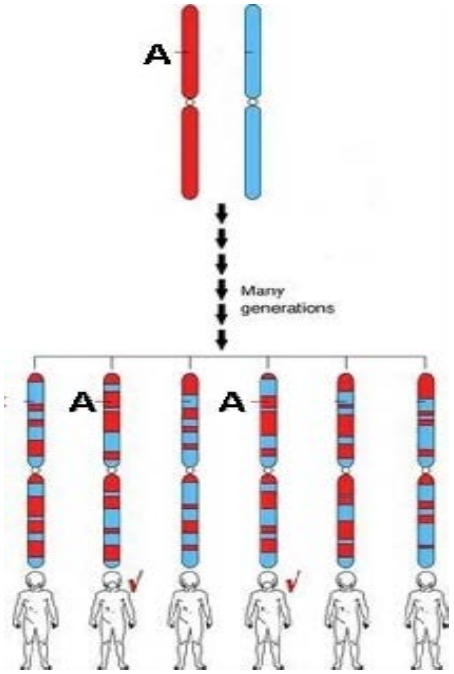


The 1000 Genomes Project is an international research consortium formed to create the most detailed and medically useful picture to date of human genetic variation. The project involves sequencing the genomes of approximately 1200 people from around the world and receives major support from the Wellcome Trust Sanger Institute in Hinxton, England, the Beijing Genomics Institute Shenzhen in China and the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH).

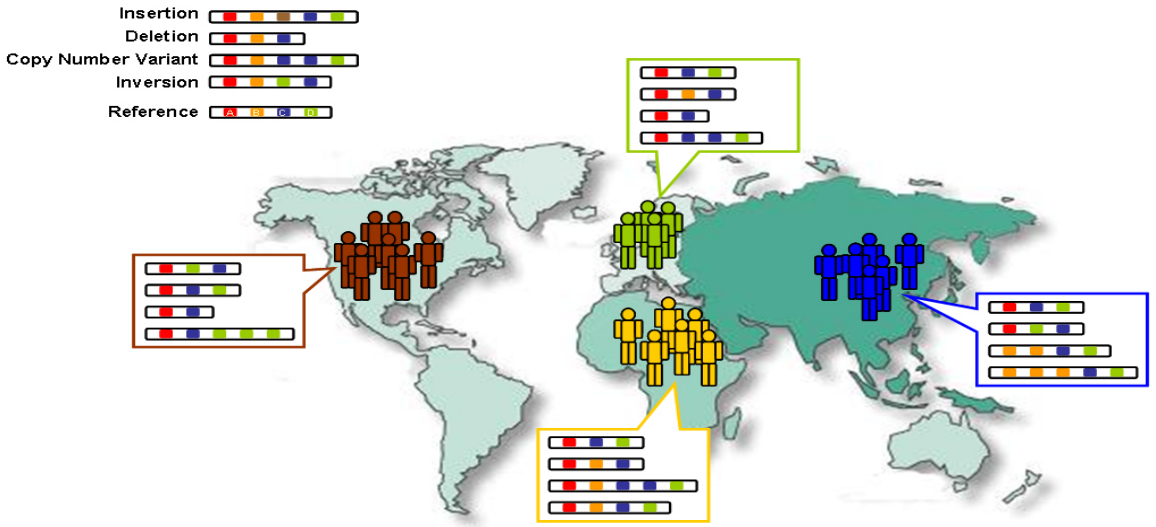
Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

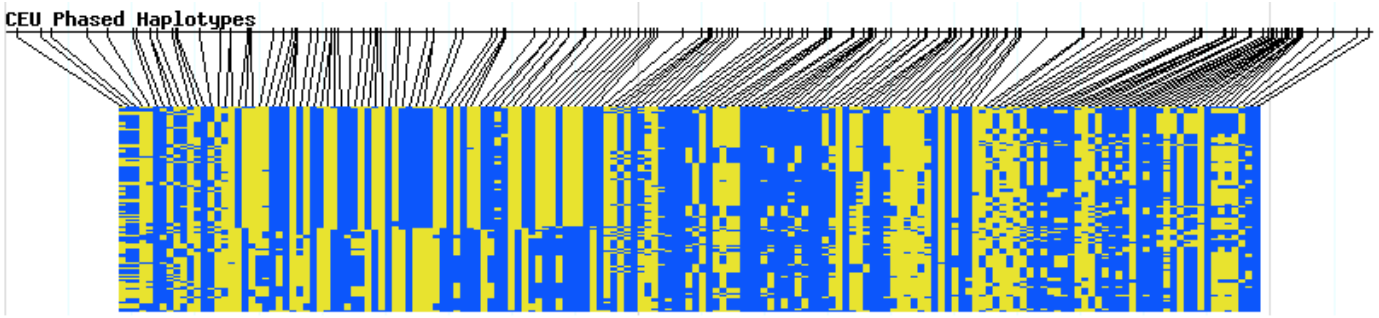
On 4 September 2008, the co-chairs of the analysis group and overall project co-chairs drafted a letter to the NIH Council about 1000 Genomes Project. This letter, available here, reviews the goals, describes the current status, and provide an update on the critical tasks the Analysis Group must accomplish in order to deliver a valuable community resource and achieve the Project's goals.

The Scientific Goals of the 1000 Genomes Project



To produce a catalog of variants that are present at 1 percent or greater frequency in the human population across most of the genome, and down to 0.5 percent or lower within genes.





A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium

Affiliations | **Contributions** | **Corresponding author**

Nature **467**, 1061–1073 (28 October 2010) | doi:10.1038/nature09534
Received 20 July 2010 | Accepted 30 September 2010 | Published online 27 October 2010

Abstract

Abstract • Introduction • Data generation, alignment and variant discovery • Power to detect variants • Genotype accuracy • Putative functional variants • Application to association studies • Mutation, recombination and natural selection • Discussion • Methods • References • Acknowledgements • Author information • Supplementary information • Comments

The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. Here we present results of the pilot phase of the project, designed to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. We undertook three projects: low-coverage whole-genome sequencing of 179 individuals from four populations; high-coverage sequencing of two mother–father–child trios; and exon-targeted sequencing of 697 individuals from seven populations. We describe the location, allele frequency and local haplotype structure of approximately 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants, most of which were previously undescribed. We show that, because we have catalogued the vast majority of common variation, over 95% of the currently accessible variants found in any individual are present in this data set. On average, each person is found to carry approximately 250 to 300 loss-of-function variants in annotated genes and 50 to 100 variants previously implicated in inherited disorders. We demonstrate how these results can be used to inform association and functional studies. From the two trios, we directly estimate the rate of *de novo* germline base substitution mutations to be approximately 10^{-8} per base pair per generation. We explore the data with regard to signatures of natural selection, and identify a marked reduction of genetic variation in the neighbourhood of genes, due to selection at linked sites. These methods and public data will support the next phase of human genetic research.

日本語要約
print
email
download pdf
download citation
order reprints
rights and permissions
share/bookmark

1000 Genomes Samples							05-Mar-10
Population	Status	Cell lines avail (all data approx)	Adult/child samples from trios avail	First set	Second set	Third set	Total
Utah residents (CEPH) with Northern and Western European ancestry (CEU)	Available - HapMap samples	At Coriell	yes	100			100
Tuscan in Italia (TSI)	Available - HapMap samples	At Coriell		100			100
British from England and Scotland (GBR)	Available	At Coriell		100			100
Finnish from Finland (FIN)	Available	At Coriell		100			100
Iberian populations in Spain (IBS)	Collecting samples	Jun 2010	yes		100		100
TOTAL European ancestry				400	100		500
Han Chinese in Beijing, China (CHB)	Available - HapMap samples	At Coriell		100			100
Japanese in Tokyo, Japan (JPT)	Available - HapMap samples	At Coriell		100			100
Han Chinese South (CHS)	Available	At Coriell	yes	100			100
Chinese Dai in Xishuangbanna (CDX)	Awaiting govt approval	late 2010	yes		100		100
Kinh in Ho Chi Minh City, Vietnam (KHV)	Awaiting govt approval	late 2010	yes		100		100
Chinese in Denver, Colorado (CHD) (pilot 3only)	Available - HapMap samples	At Coriell		0	0		0
TOTAL East Asian ancestry				300	200		500
Yoruba in Ibadan, Nigeria (YRI)	Available - HapMap samples	At Coriell	yes	100			100
Luhya in Webuye, Kenya (LWK)	Available - HapMap samples	At Coriell		100			100
Gambian in Western Division, The Gambia (GWD)	Collecting samples	late 2010	yes		100		100
Ghanaian in Navrongo, Ghana (GHN)	Final IRB approval expected soon	late 2010	yes		100		100
Malawian in Blantyre, Malawi (MAB)	Discussing issues for participation	?	yes		100		100
TOTAL West African ancestry				200	300		500
African Ancestry in Southwest US (ASW)	Available - HapMap samples	At Coriell	yes	61			61
African American in Jackson, MS (AJM)	IRB approval received, recruit soon	late 2010	yes		60		60
African Caribbean in Barbados (ACB)	Collecting samples	late 2010	yes		79		79
Mexican Ancestry in Los Angeles, CA (MXL)	Available - HapMap samples	At Coriell	yes	70			70
Puerto Rican in Puerto Rico (PUR)	Available	Jun 2010	yes	70			70
Colombian in Medellin, Colombia (CLM)	Collecting samples	Jun 2010	yes		70		70
Peruvian in Lima, Peru (PEL)	Collecting samples	late 2010	yes		70		70
TOTAL Americas				201	229		500
Ahom in the State of Assam, India	Awaiting govt & IRB approval	2011	yes			100	100
Kayasths in Calcutta, India	Awaiting govt & IRB approval	2011	yes			100	100
Reddy in Hyderabad, India	Awaiting govt & IRB approval	2011	yes			100	100
Marathi in Bombay, India	Awaiting govt & IRB approval	2011	yes			100	100
Punjabi in Lahore, Pakistan	Awaiting IRB approval	late 2010	yes			100	100
TOTAL South Asian ancestry						500	500
TOTAL				1101	899	500	2500

ARTICLE

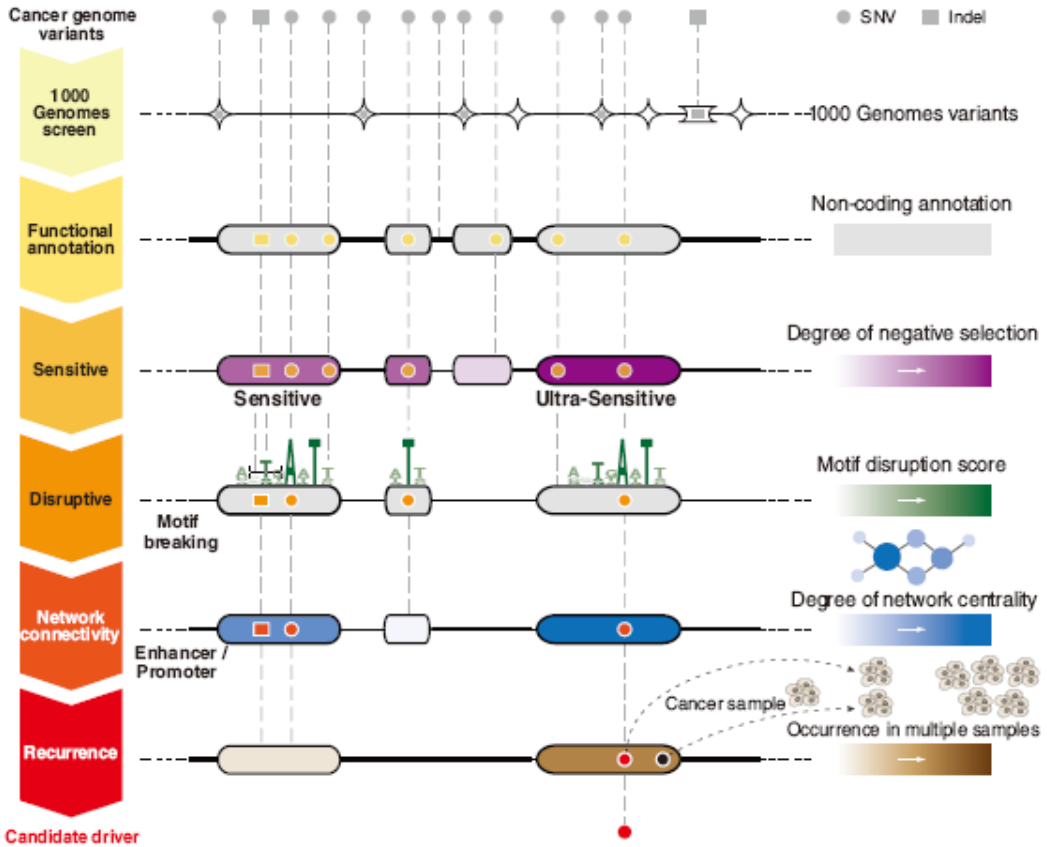
doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

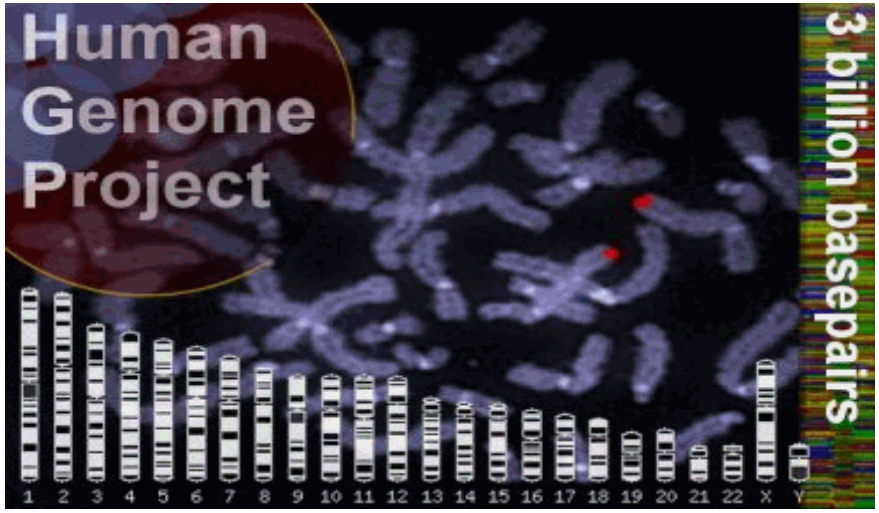
By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

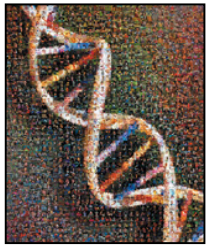
Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics



Prioritization of candidate noncoding cancer drivers based on patterns of selection. (Step 1) Filter somatic variants to exclude 1000 Genomes polymorphisms; (2) retain variants in noncoding annotations; (3) retain those in "sensitive" regions; (4) prioritize those disrupting a transcription-factor binding motif and (5) residing near the center of a biological network; (6) prioritize ones in annotation blocks mutated in multiple cancer samples.

How much does it cost for a Human Genome?

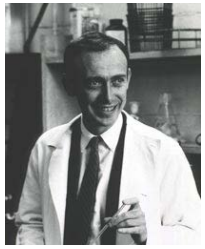




Human Reference Genome

April, 2003

>10 years to finish
USD 3 billion



James Watson's Personal Genome

June, 2007

1 year
USD 2 million



Craig Venter's Personal Genome

September, 2007

1 year
USD 1 million



YH Genome

November, 2008

1 year
~USD 0.5 million



Personal Genome

When will it available?
How long will it take?
How much?

Sequenced Human Genomes

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000‖‖
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000‖‖
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

USD 10000 per Genome in 2015?

"The scaffold has been handed down to us from our ancestors, and through it we are connected to all other life on earth."

- Svante Pääbo

Revolution Through Competition.

TAKE ACTION

ARCHON X PRIZE FOR GENOMICS

- Introduction
- Why Genomics?
- PRIZE Overview
- Why Whole Genome Sequencing?
- The Promise of Personalized Medicine
- Frequently Asked Questions
- Competition Guidelines [PDF]
- Register to Compete [PDF]



A $10 MILLION PRIZE FOR THE FIRST TEAM TO SUCCESSFULLY SEQUENCE 100 HUMAN GENOMES IN 10 DAYS

ARCHON X PRIZE FOR GENOMICS

Scientists know that a map of our genome holds boundless potential, ranging from identifying our susceptibility to disease to discovering cures for cancer. But since 1953, when James Watson and Francis Crick concluded that DNA contained the "stuff of life," only a handful of human genomes have been mapped. In fact, it still takes many months and millions of dollars to sequence a single genome.

Understanding our genomes may help delay or even prevent disease. For those suffering from genetic illnesses, personal genetic information can determine which medicines will drive their disease into remission without negative side-effects.

The Archon X PRIZE for Genomics challenges scientists and engineers to create better, cheaper and faster ways to sequence genomes. The knowledge gained by compiling and comparing a library of human genomes will create a new era of preventive and personalized medicine — and transform medical care from reactive to proactive.

The X PRIZE Foundation and scientists the world over dream of the day when we fully understand the human genetic blueprint — enabling us to make informed decisions about our own health and create a brighter future for generations to come.

Please join the X PRIZE Foundation in our challenge to create a breakthrough in genomics that will benefit all of humanity.

Join the Revolution

"Many thought that the unbelievable feat in 2001 of deciphering the human genome was the end of the story. Not so! We now firmly believe that it was only the beginning. The true challenge now appears to be in deciphering individual variations within the human genome. It is there that the true importance of genomics resides. Only by unearthing the blueprint details of each and every individual, will we be able to fully fathom health and disease, mental and bodily faculties, and the true secrets of human biology. The Archon X PRIZE for Genomics will greatly enhance scientists' ambition to develop methodologies for individual genomic "blue-printing". I believe the PRIZE will bring about bona-fide successes within 5 years, much earlier than otherwise feasible. It will thereby create the real genomic revolution."

Prof. Doron Lancet
Head, Crown Human Genome Center
Weizmann Institute of Science
Archon X PRIZE for Genomics Scientific Advisory Board

Knome

Home | Solutions | Technology | Company | Contact

Sequencing & interpretation: $4,998 per whole genome

Minimum order of 10 genomes

Learn More



Rare Disease

Identify the variants that underlie rare disease.

Common Disease

Unravel the complex basis of common disease.

Cancer

Find and understand the drivers of tumor growth.

Drug Response

Pinpoint the genetic variants that govern drug response.

Complete genomics

Revolutionizing human genome discovery



Human Genome Sequencing & Analysis Service

Dedicated to **complete human genome sequencing and analysis** provided as an innovative, **end-to-end, outsourced service model**, Complete Genomics enables researchers to conduct **large-scale complete human genome studies**.

By **optimizing our sequencing platform for human DNA**, we are able to achieve **accuracy levels of 99.999%** at a total cost that is significantly less than the total cost of purchasing and using commercially available DNA sequencing instruments.

» read more

Receive Research-Ready Genomic Data

We offer our customers an end-to-end, outsourced solution that delivers **research-ready genomic data**.

Our CGA™ Service provides reliable access to assembled and **annotated sequence data** and our analytical tools enable our customers to **rapidly analyze and compare genomic data**.

» read more

WEBINAR

Pedigree Genome Sequencing



presented by:
Jared Roach, Ph.D., M.D.,
Institute for Systems Biology

View

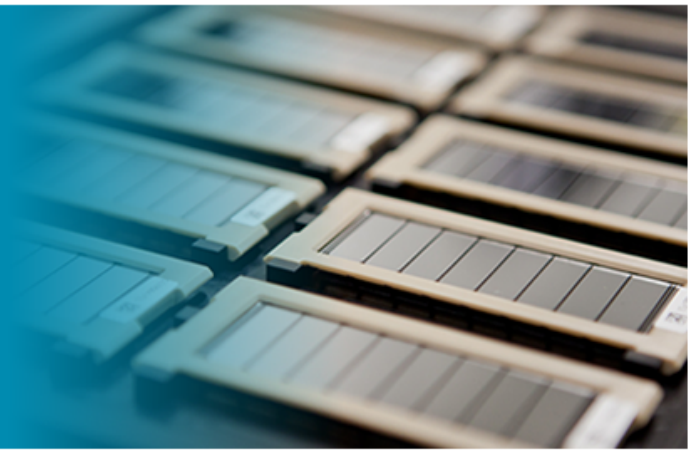
News and Events

» **Complete Genomics Announces Pricing of Initial Public Offering**

» **Complete Genomics Expands CGA Tools Software Suite**

» **Complete Genomics to Sequence**

COLLABORATOR PUBLICATION: GENOME SEQUENCING IDENTIFIES MAJOR CAUSES OF SEVERE INTELLECTUAL DISABILITY



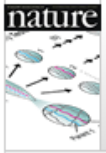
Accurate Whole Human Genome Sequencing & Analysis

Complete Genomics is a leader in whole human genome sequencing based in Mountain View, California. Using its proprietary sequencing instruments, chemistry, and software, the company has sequenced more than 20,000 whole human genomes over the past three years. The company's mission is to improve human health by providing researchers and clinicians with the core technology and commercial systems to understand, prevent, diagnose, and treat diseases and conditions.

nature

Genome sequencing identifies major causes of severe intellectual disability

nature



Read the ground breaking Nature paper on LFR Technology

Read More

News

Complete Genomics Workshop at ASHG 2014: Thought Leaders Address Challenges, Breakthroughs, and Innovations in Clinical Whole Genome Sequencing

Events

Personalized Medicine World Conference, Silicon Valley 2015

ACMG Annual Clinical Genetics Meeting 2015

Personalized Medicine World Conference, United Kingdom 2015

http://www.macrogen.com/

MACROGEN

Company | R & D | Service | IR | News | CS Center

NGS | Sequencing | Microarray | Mouse | Oligo | Bio Informatics

Welcome to the era of X-Genome, the Xpressway to Genome

The $1,000 genome era has officially begun as Macrogen becomes one of the first in the world to acquire new HiSeq X Ten Sequencing System.



What is X-Genome?

X-Genome is Macrogen's human whole genome sequencing product line based on Illumina's HiSeq X Ten Sequencing System.

read more »

X-Genome Applications

Macrogen is going to fully utilize Illumina's new groundbreaking technology, HiSeq X to take on various scientific and medical challenges.

read more »

FAQ

GO

Media Release

The $1,000 genome era has officially begun as Macrogen becomes one of the first in the world to acquire new HiSeq X Ten Sequencing System.

read more »

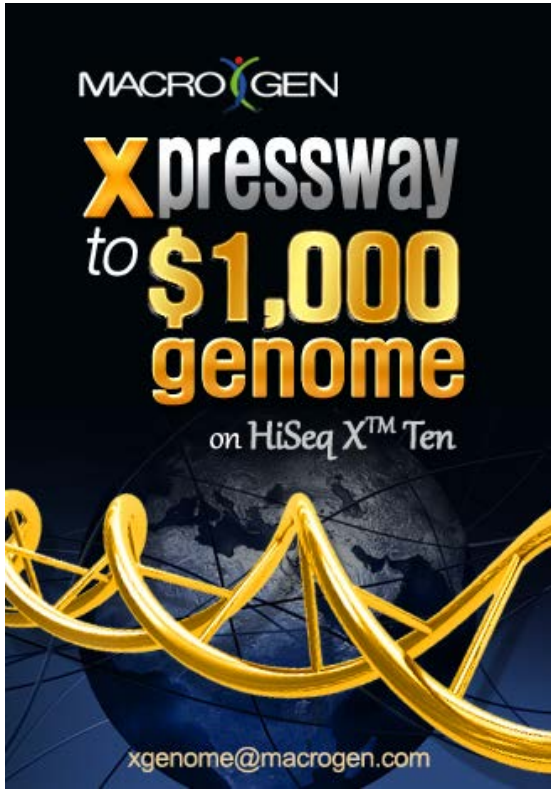
X-Genome Product Description

We are building a comprehensive X-Genome product line to accommodate your research and clinical needs.

read more »

Contact us

GO

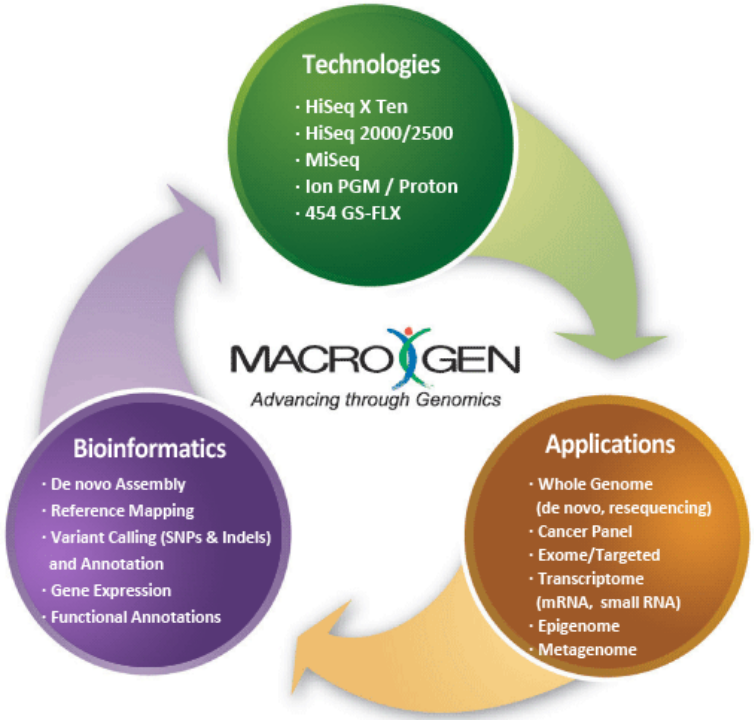


Chairman

The first message to man-kind of the 21 century was to complete the human genome map



Jeongsun Seo
Macrogen's Chairman



Population power. Extreme throughput. $1,000 human genome.

The HiSeq X Ten is a set of ten ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.



Table 1: HiSeq X Ten Performance Parameters*

	Dual Flow Cell	Single Flow Cell
Output/Run	1.6–1.8 Tb	800–900 Gb
Reads Passing Filter†	5.3–6 billion	2.6–3 billion
Supported Read Length	2 × 150	
Run Time	< 3 days	
Quality	≥ 75% of bases above Q30 at 2 × 150 bp	

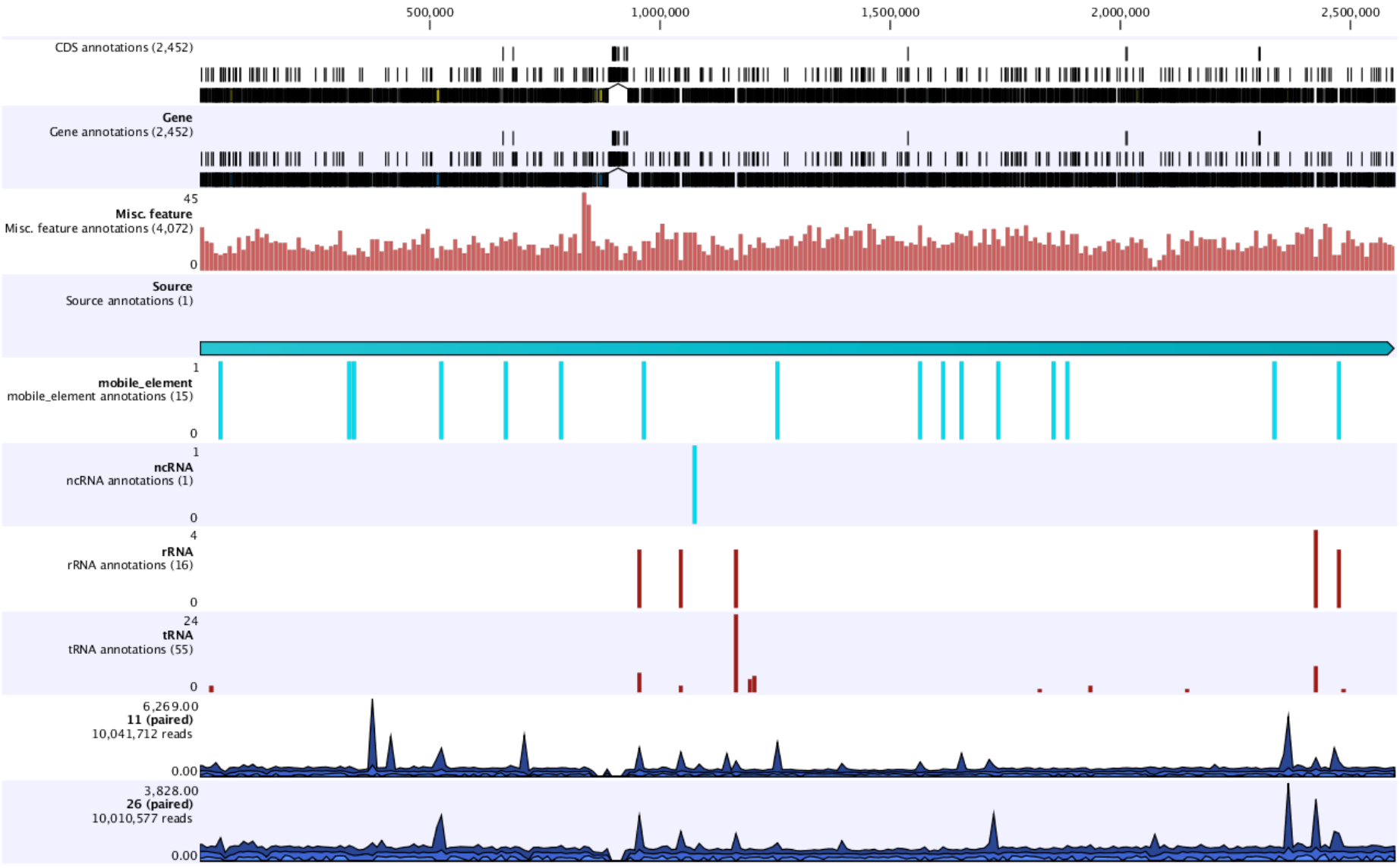
* Specifications based on Illumina PhiX control library at supported cluster densities (between 1,255–1,412 K clusters/mm²). Supported library preparation kit includes TruSeq Nano DNA HT kit with 350 bp target insert size and HiSeq X HD reagents. HiSeq X was designed and optimized for human whole-genome sequencing; other applications and species are not supported.
† Single-end reads.

Figure 2: The HiSeq X Ten Enables Sequencing of Tens of Thousands of Human Genomes per Year

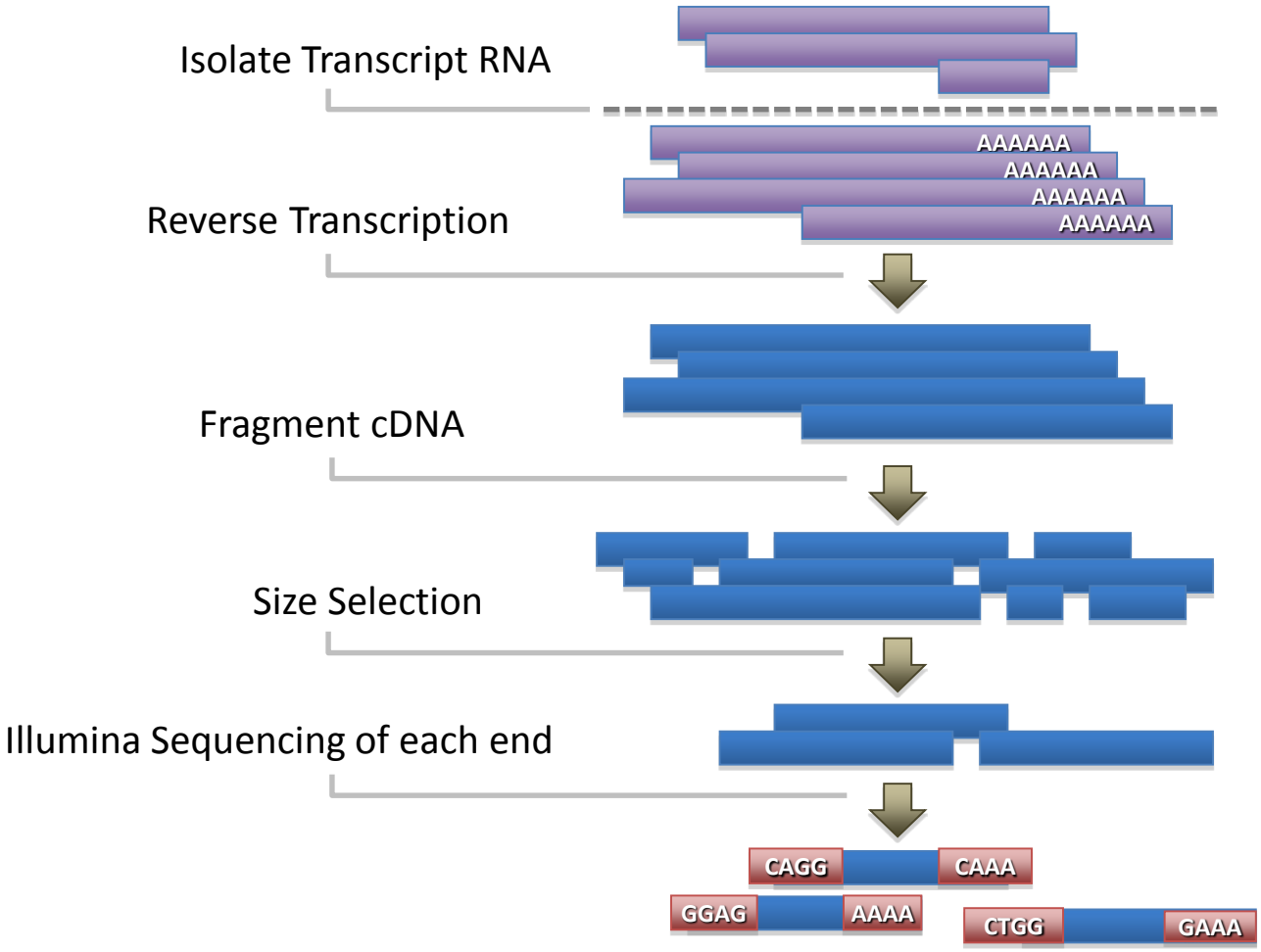


When operating in parallel, the 10 instruments of a HiSeq X Ten generate a staggering level of throughput, capable of sequencing tens of thousands of genomes per year.

Bacteria Re-sequencing



Transcriptome Sequencing



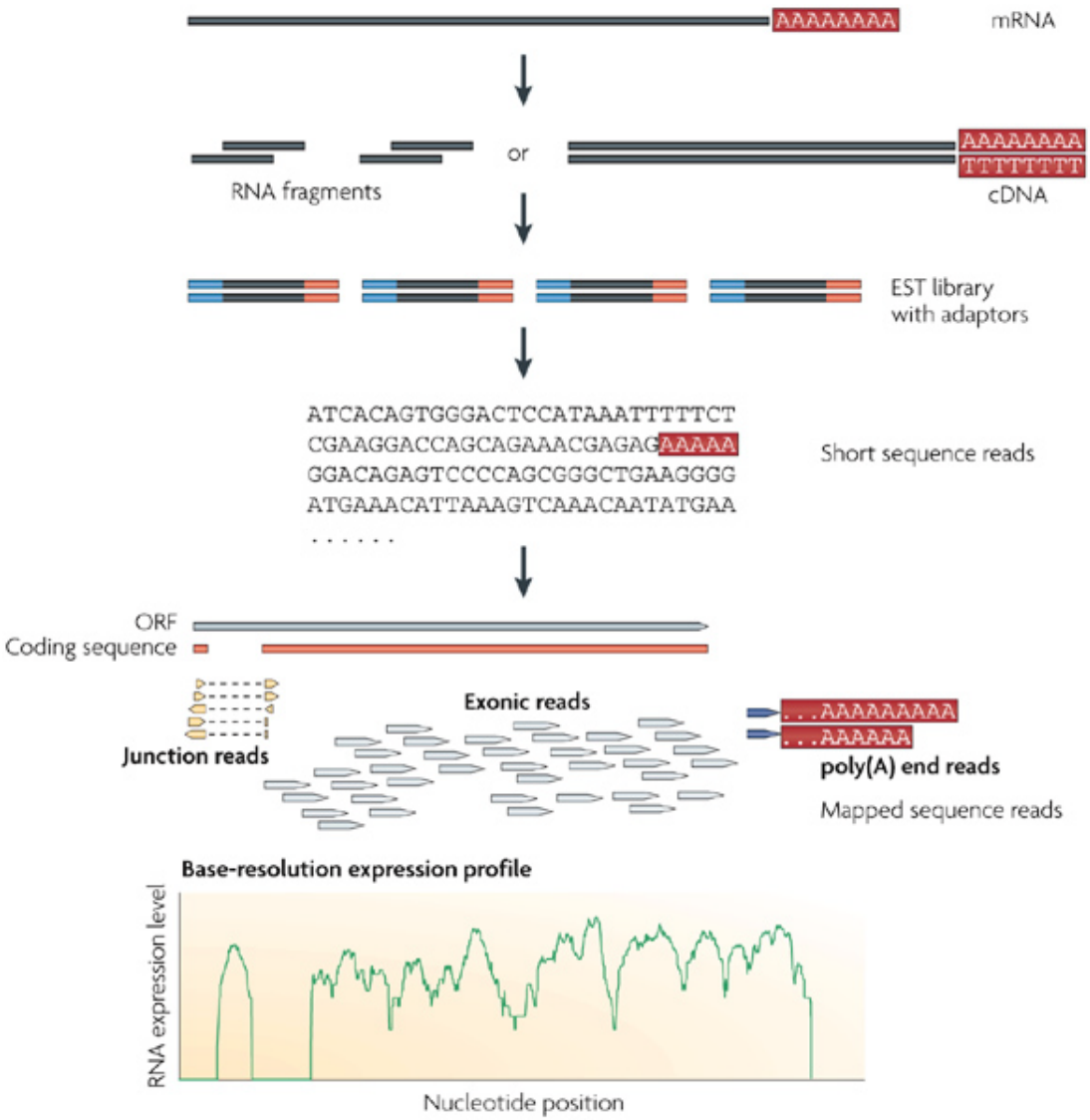
XX,000,000 reads

*based on Illumina approach

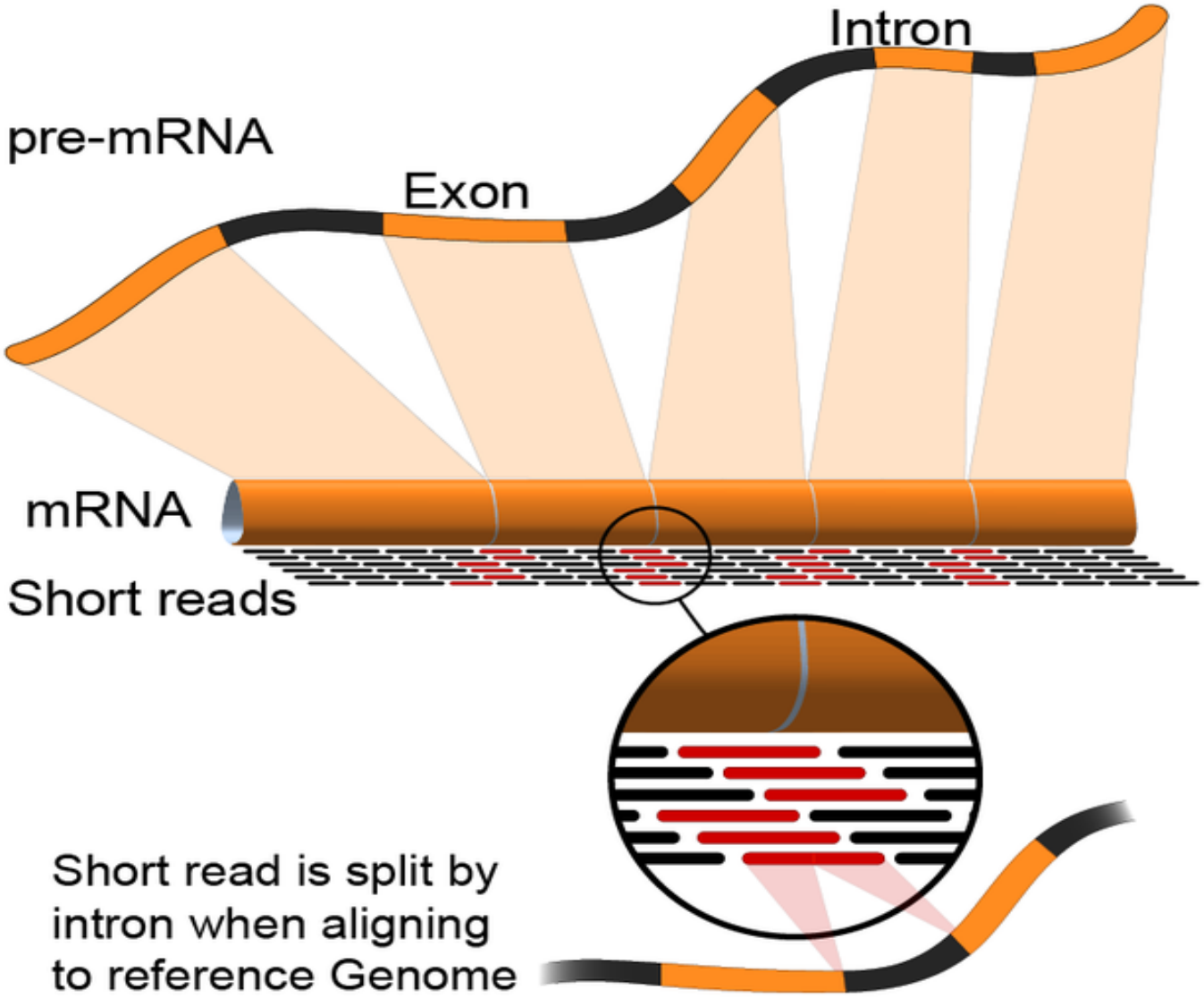
**strand-specific RNA-seq protocols exist for both Illumina and SOLiD

Slide complements of Andrew McPherson

Overview of RNA-Seq



Transcriptome Sequencing with Reference



Analysis Strategies: Reference Sequence Alignment *vs* De novo Assembly

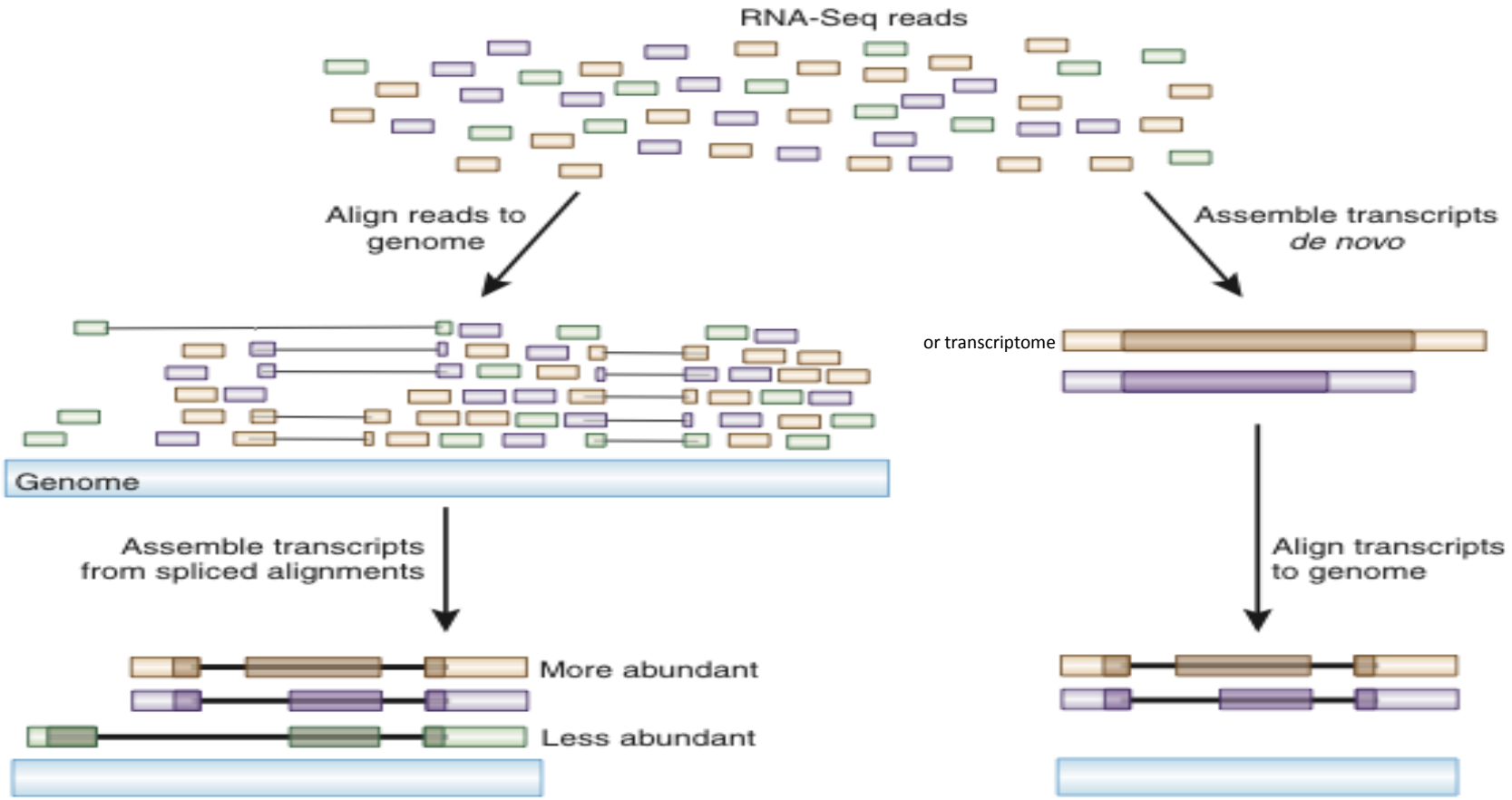
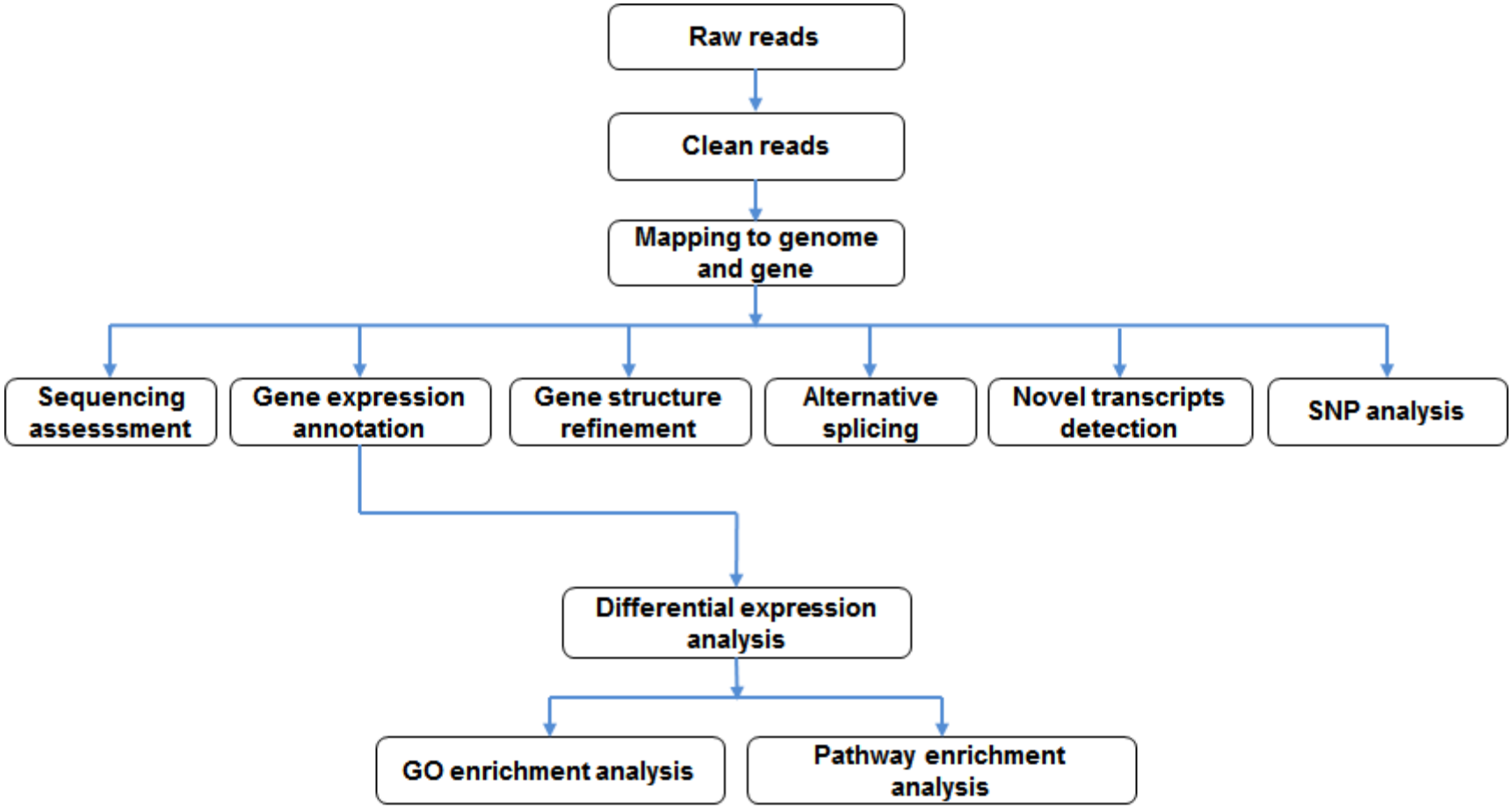


Image from Haas & Zody, 2010

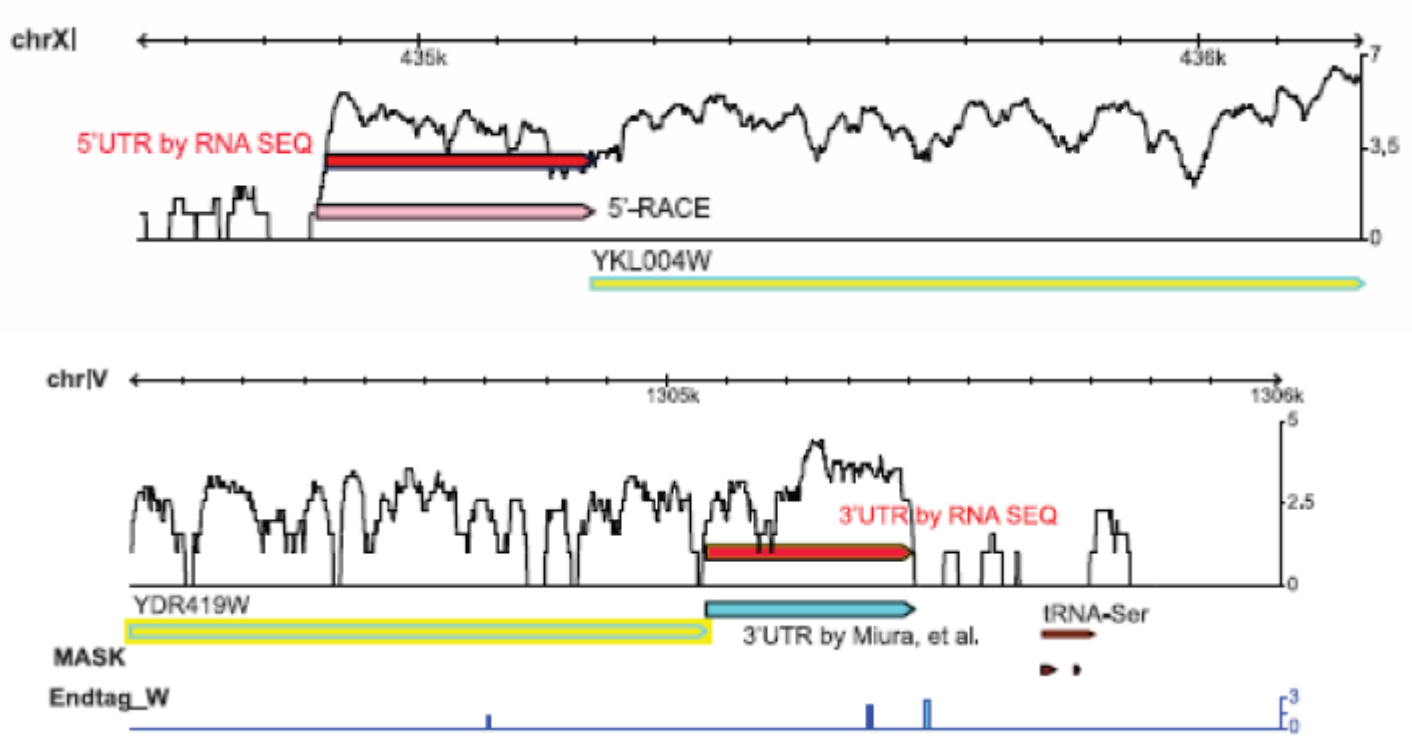
*Assembly is the only option when working with a creature with no genome sequence, alignment of contigs may be to ESTs, cDNAs etc

Bioinformatics Analysis Pipeline for Transcritpome Sequencing with Reference



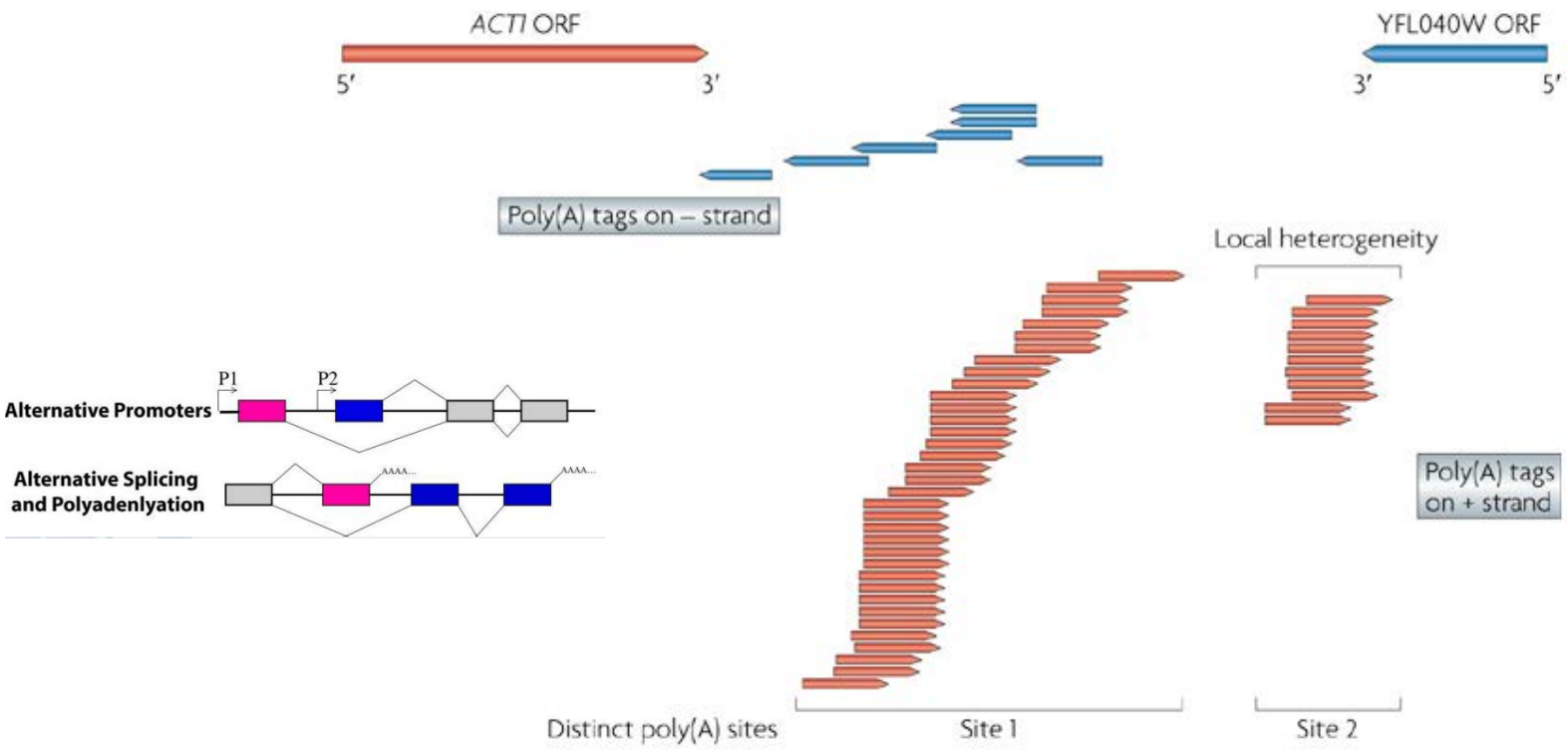
Gene structure refinement

UTRs by RNA-Seq in yeast



(Nagalakshmi, U. et al.,2008)

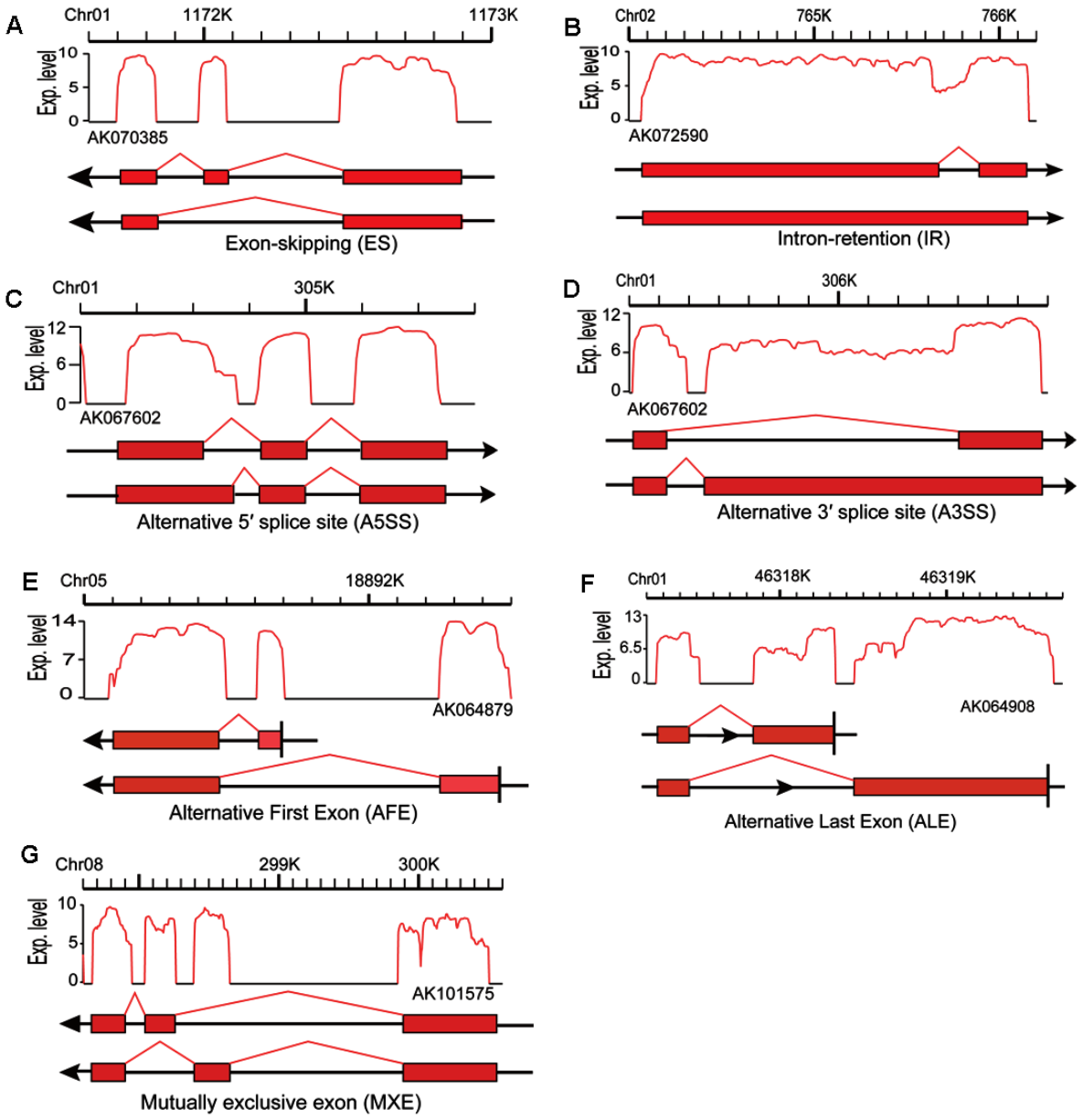
Poly(A) tags from RNA-Seq



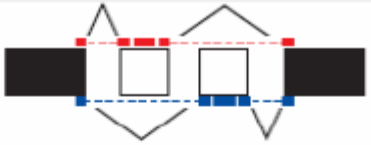
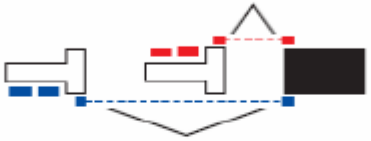
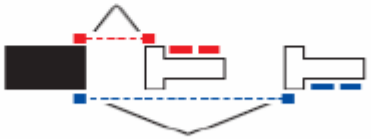
A region containing two overlapping transcripts (ACT1, from the actin gene, and YFL040W, an uncharacterized ORF) from the Saccharomyces cerevisiae genome is shown. Arrows point to transcription direction. The poly(A) tags from RNA-Seq experiments are shown below these transcripts, with arrows indicating transcription direction. The precise location of each locus identified by poly(A) tags reveals the heterogeneity in poly(A) sites, for example, ACT1 has two big clusters, both with a few bases of local heterogeneity. The transcription direction revealed by poly(A) tags also helps to resolve 3'-end overlapping transcribed regions

Nature Reviews Genetics ***10****, 57-63*

Alternative Splicing (AS)

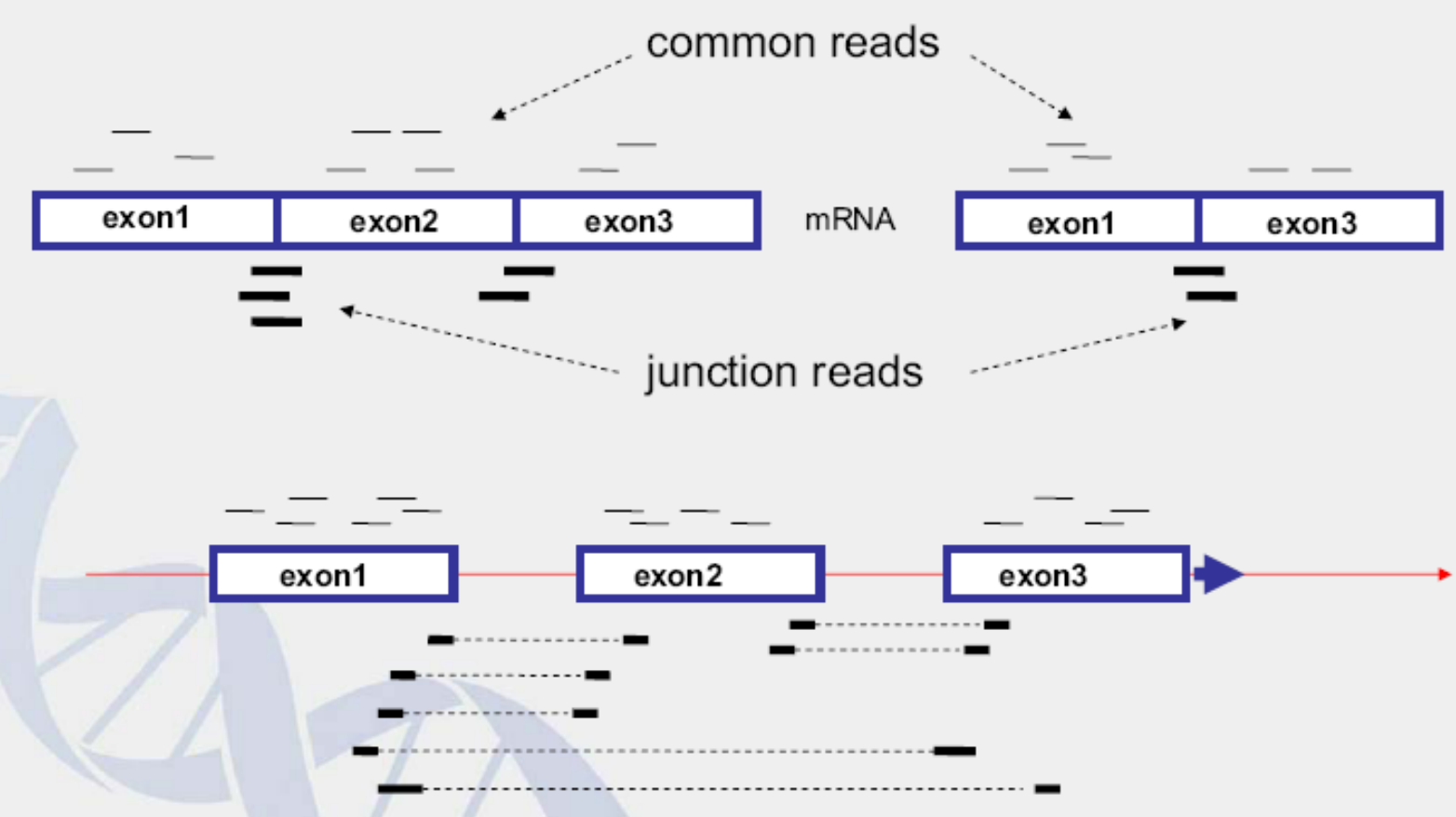


92%~94% of human multi-exon gene undergo Alternative Splicing (AS)

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)
Mutually exclusive exon (MXE)		4	4
Alternative first exon (AFE)		14	13
Alternative last exon (ALE)		9	8
Tandem 3′ UTRs		7	7



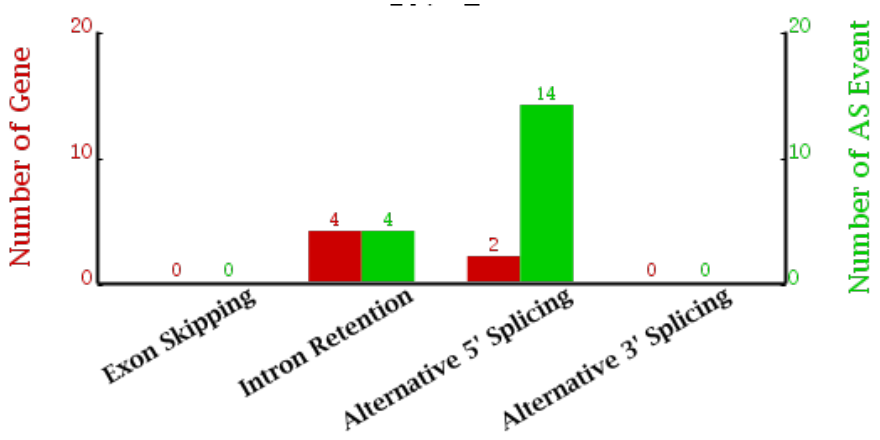
Exon junction reads



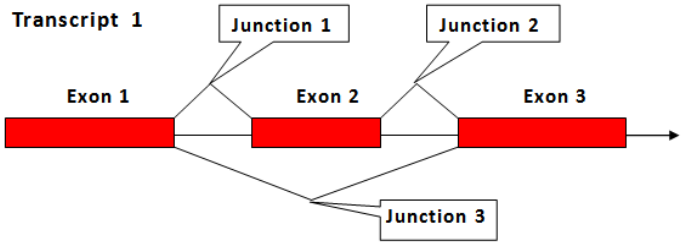
SOAP Short Oligonucleotide Analysis Package

Output of SOAPals

D	E	F	G	H
Name	strand	retained intron	Gene	Location
	+	34279-34388		
CAMK family protein kinase	-	4588-4721	3,745 to 5,333	complement(join(4722..5333,3745..4587))
	-	4588-4721		
	+	99395-99499		
	+	99395-99499		
CAMK family protein kinase	-	56818-56916	56,024 to 57,470	complement(join(57457..57470,56917..57457,56069..56817,56024..56069))
	-	56818-56916		
	-	56818-56916		
	-	56818-56916		
	-	23850-23917		
	-	72375-72446		
nuclear lim interactor-interacting factor	-	8624-8699	7,991 to 9,110	complement(join(8700..9110,8066..8623,7991..8066))
	-	8624-8699		
	-	8624-8699		
	-	8624-8699		
	-	8624-8699		
ubiquitin	+	26766-26993	26,229 to 27,193	join(26229..26765,26994..27146,27146..27193)
	+	26766-26993		
	-	9497-9589		
	-	78748-78814		
	-	78748-78814		



Splice Junction Database

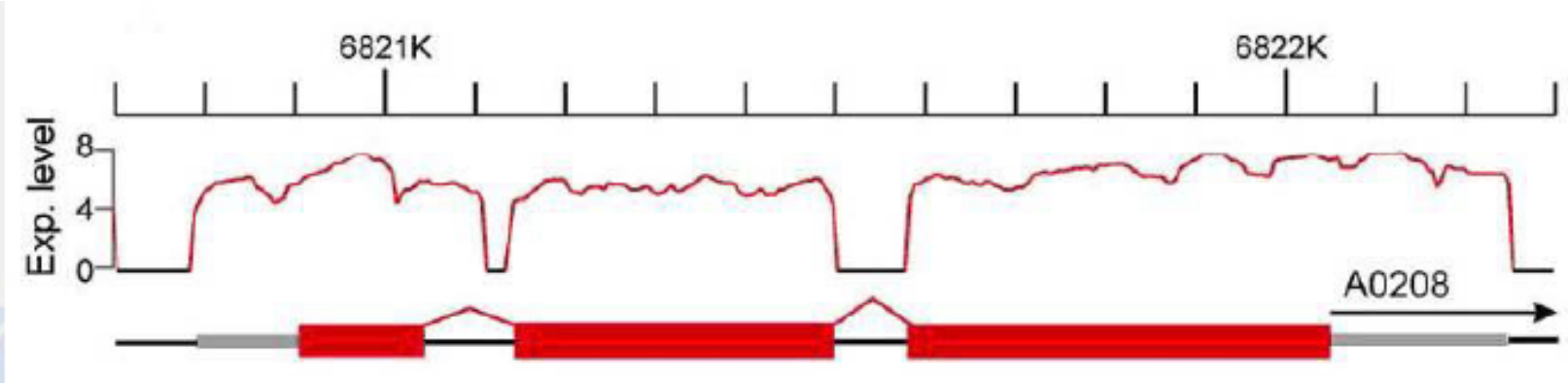


Species: Oryza Sativa
Chromosome: 1
All the Genes

```
SELECT * FROM oryzasativa_junction where GeneID=4326813;
```

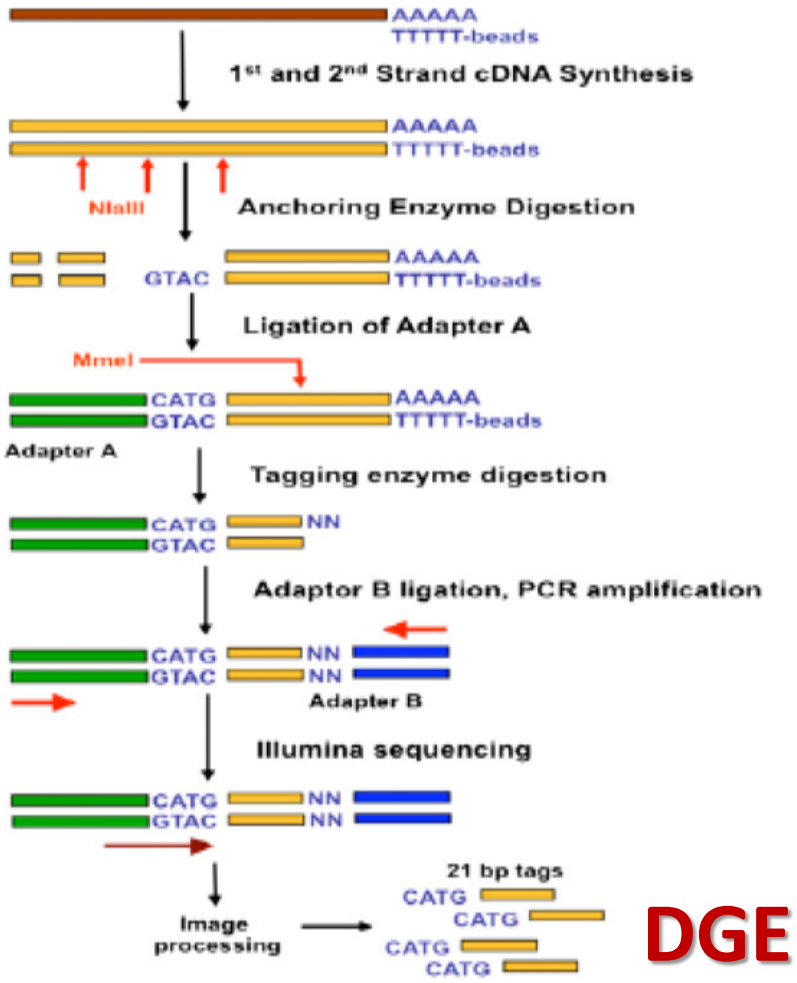
JunctionID	GeneID	exon3P	exon5P	splicejunc
490591	4326813	1	2	TTGGCGACCCCTGCCTCAACCCATCCCCCGTTAAGGGGAGCAAAATGCTCAAACCGGAAAAATGGCACAC
490592	4326813	2	3	GTTTCCGTAAAGCCCTAAAATTCATTGTCTTAGGGGGTGTGGATCCCACTATTCGAGCTGAAGTTTGGGA
490593	4326813	3	4	TCAGAGTATAGGAGGAAACTAAGAGCTGTTAGAAGGGAAAAATATCAAATTTTAGTTAGACAGTGCCAGA
490594	4326813	4	5	GCTGTAGTAAATCAGCTGAACTAGTTGGTTTCAACGATTACAATGACATGGGTGAACCAAGATATGACAC
490595	4326813	5	6	AGTCTCTGAATGGCTTTGGACACTGCACCGAATTGTGGTTGATGTGGTTAGAACGGATAGTCATCTTGAT
490596	4326813	6	7	CCTTTTGGTGCTTTGAGATGCTACTGAGAAGGATGCGTGAAAATTTCCAGATGGAGGGACCAACAGGAGT
490597	4326813	7	8	TTAGCATGTGGGAGGTGTGGCATTTGAAGACACATATGATGTGGGCTGCTGACTTTAATGAAGATGTGAT
490598	4326813	8	9	ATGATCTGCCTATATTCTGCGTAGCAGCAATCCTTATGTTCAATGATAATATGCTGAAGATCAATGTAAA
490599	4326813	9	10	TGGCTATCAAGCTGAGGAAGAAGTATATTTACAAGGGAATCTGTACATATGCACAGTATACTTGAAAACA
490600	4326813	1	3	TTGGCGACCCCTGCCTCAACCCATCCCCCGTTAAGGGTGTGGATCCCACTATTCGAGCTGAAGTTTGGGA
490601	4326813	2	4	GTTTCCGTAAAGCCCTAAAATTCATTGTCTTAGGGGGAAAAATATCAAATTTTAGTTAGACAGTGCCAGA
490602	4326813	3	5	TCAGAGTATAGGAGGAAACTAAGAGCTGTTAGAAGGATTACAATGACATGGGTGAACCAAGATATGACAC
490603	4326813	4	6	GCTGTAGTAAATCAGCTGAACTAGTTGGTTTCAACTGGTTGATGTGGTTAGAACGGATAGTCATCTTGAT
490604	4326813	5	7	AGTCTCTGAATGGCTTTGGACACTGCACCGAATTGCGTGAAAATTTCCAGATGGAGGGACCAACAGGAGT
490605	4326813	6	8	CCTTTTGGTGCTTTGAGATGCTACTGAGAAGGATGATGATGTGGGCTGCTGACTTTAATGAAGATGTGAT
490606	4326813	7	9	TTAGCATGTGGGAGGTGTGGCATTTGAAGACACATATGTTCAATGATAATATGCTGAAGATCAATGTAAA
490607	4326813	8	10	ATGATCTGCCTATATTCTGCGTAGCAGCAATCCTTGGAATCTGTACATATGCACAGTATACTTGAAAACA
490608	4326813	1	4	TTGGCGACCCCTGCCTCAACCCATCCCCCGTTAAGGGAAAAATATCAAATTTTAGTTAGACAGTGCCAGA
490609	4326813	2	5	GTTTCCGTAAAGCCCTAAAATTCATTGTCTTAGGGGATTACAATGACATGGGTGAACCAAGATATGACAC
490610	4326813	3	6	TCAGAGTATAGGAGGAAACTAAGAGCTGTTAGAAGTGGTTGATGTGGTTAGAACGGATAGTCATCTTGAT
490611	4326813	4	7	GCTGTAGTAAATCAGCTGAACTAGTTGGTTTCAACCGTGAAAATTTCCAGATGGAGGGACCAACAGGAGT

Novel Transcripts

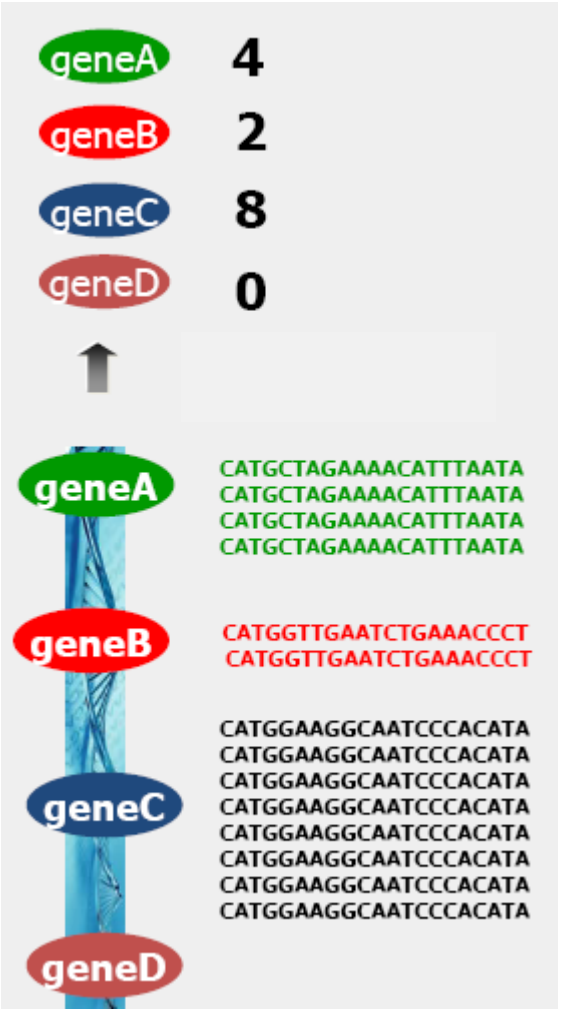


Novel transcripts can be found by high throughput sequencing since present databases may be incomplete. Gene models found in intergenic regions (200 bp away from upstream or downstream genes) were thought to be candidate of novel transcripts

Digital Gene Expression Profiling

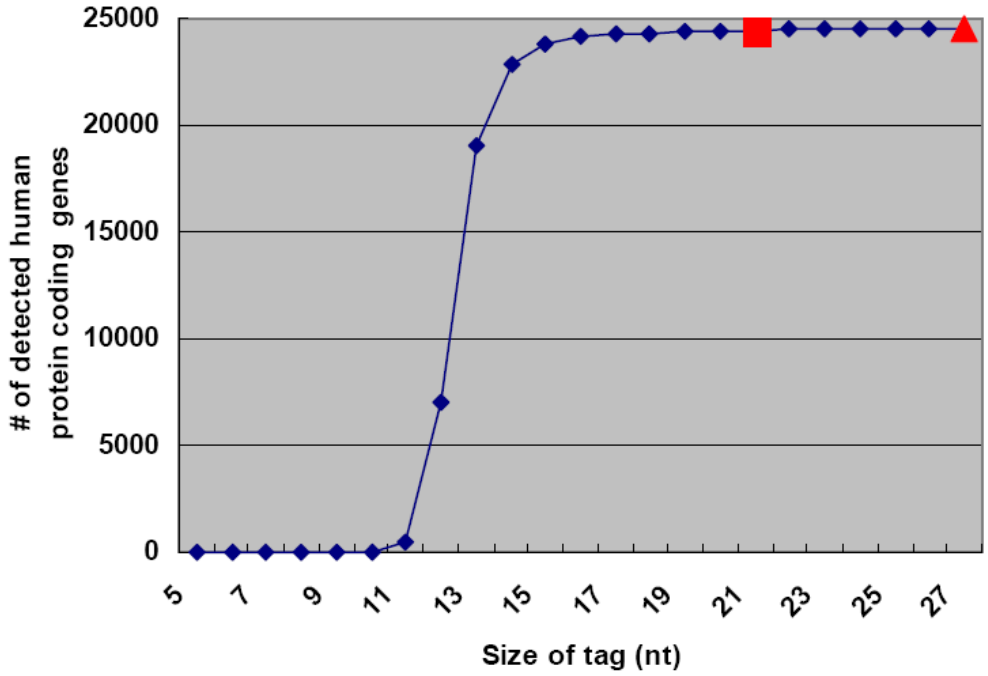


DGE



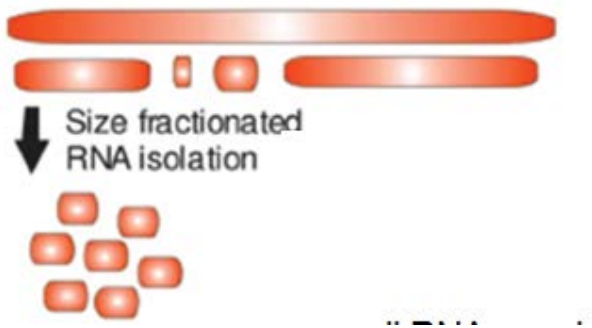
Tag length *vs* Complexity

Tag length (*n* base pairs)	Complexity[a]	Tag uniqueness probability[b]
14	1,048,576	0.00%
15	4,194,304	0.08%
16	16,777,216	16.73%
17	67,108,864	63.95%
18	268,435,456	89.43%
19	1,073,741,824	97.24%
20	4,294,967,296	99.30%
21	17,179,869,184	99.83%

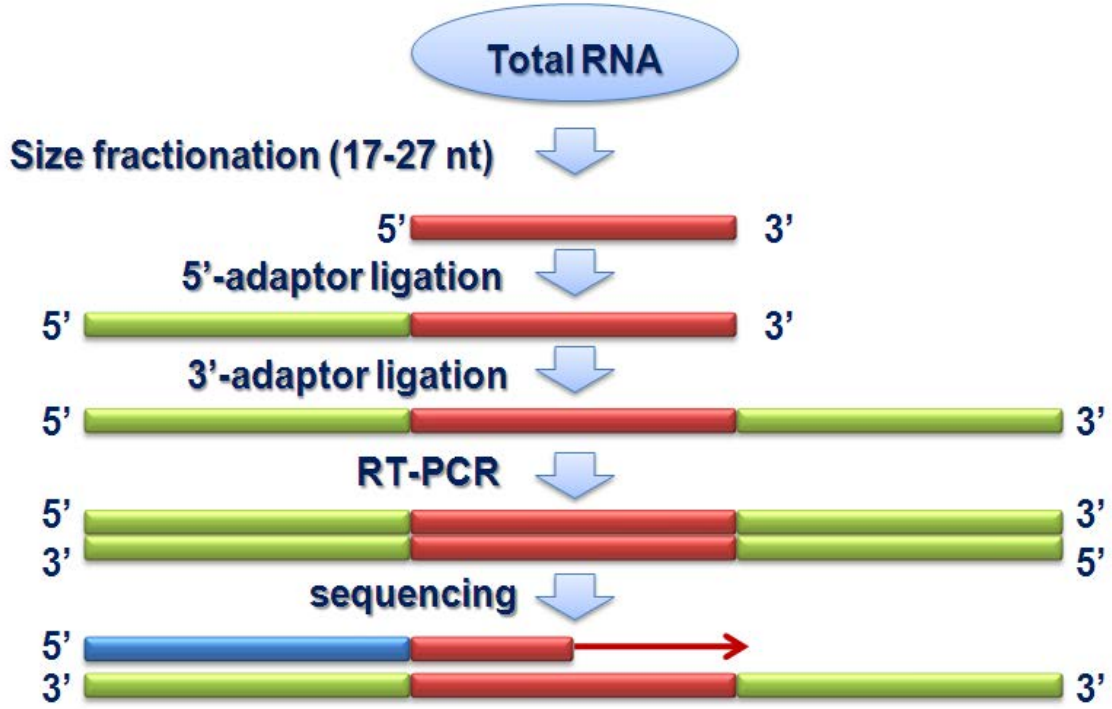


Small RNA Sequencing





Illumina SOLEXA sequencing by synthesis




```
@HWI-EAS82_3_FC204V1AAXX:6:1:886:345
AGAGTTCTACAGTCCGGACGATCTCGTATGCCGTC
+HWI-EAS82_3_FC204V1AAXX:6:1:886:345
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
```

40,000,000 reads of 35 bp long

Small RNA Sequencing

Workflow

Clean up

↓

Clustering

↓

ncRNA matching

(Rfam V10)

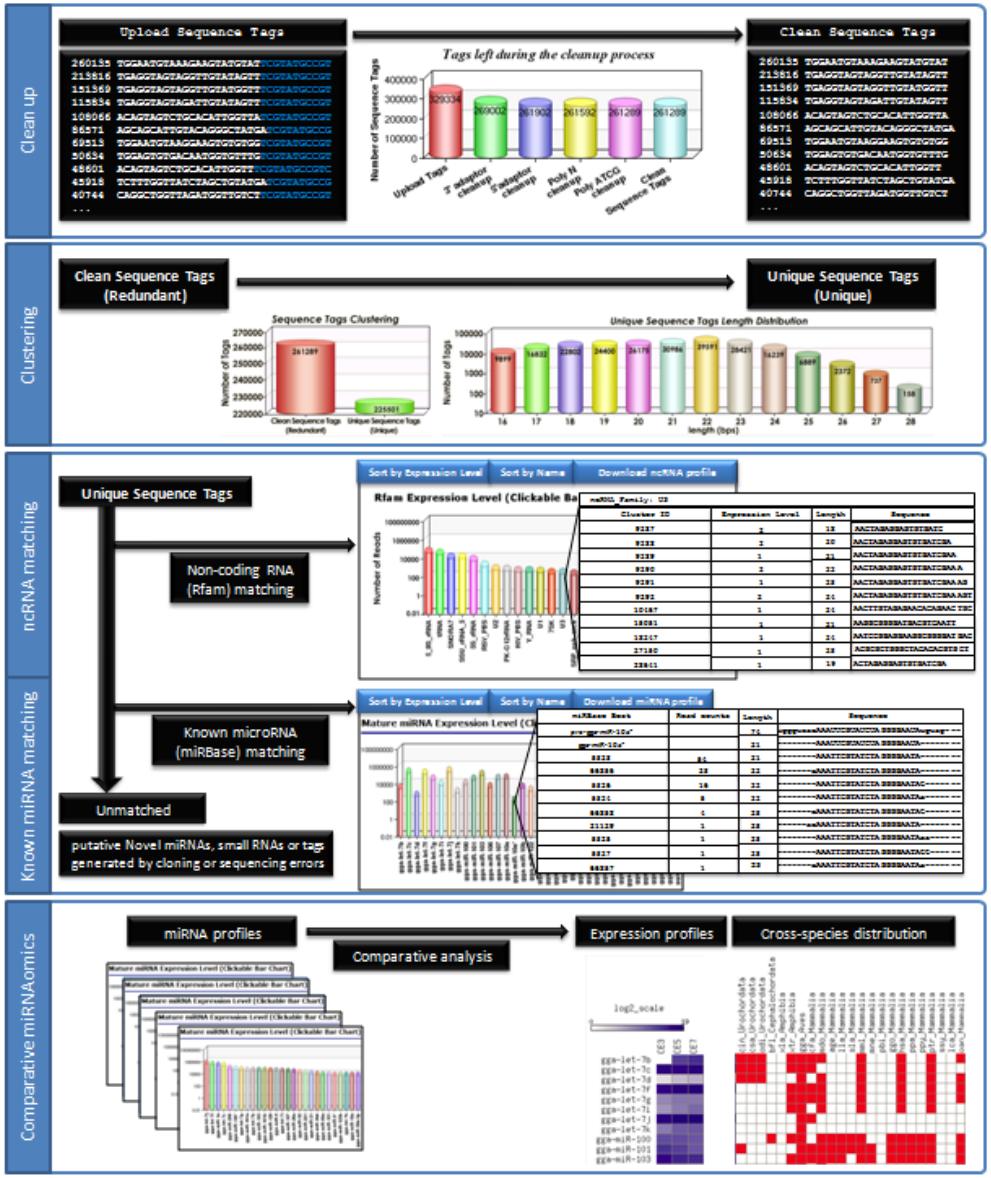
↓

Known miRNA Matching

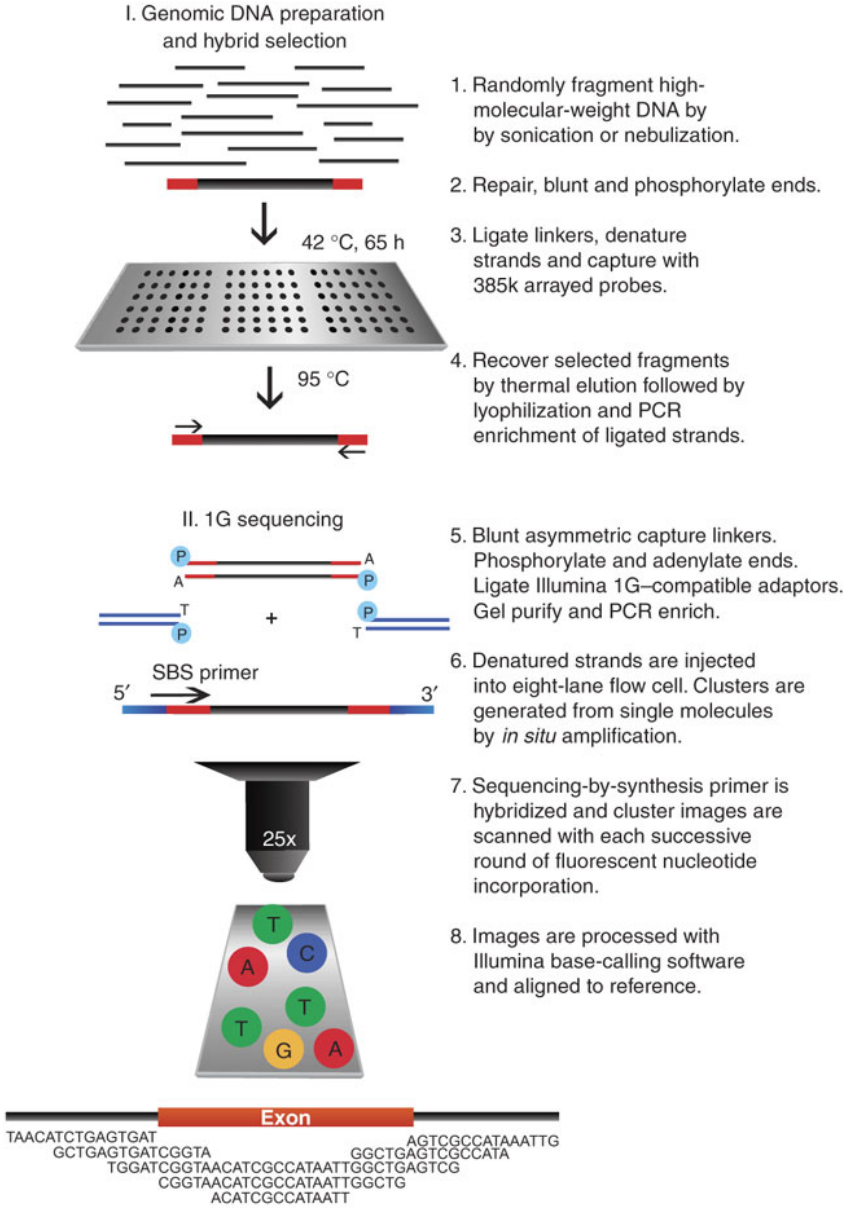
(miRBase V16)

↓

Comparative miRNAomics



Capture Sequencing: Human Exome Sequencing

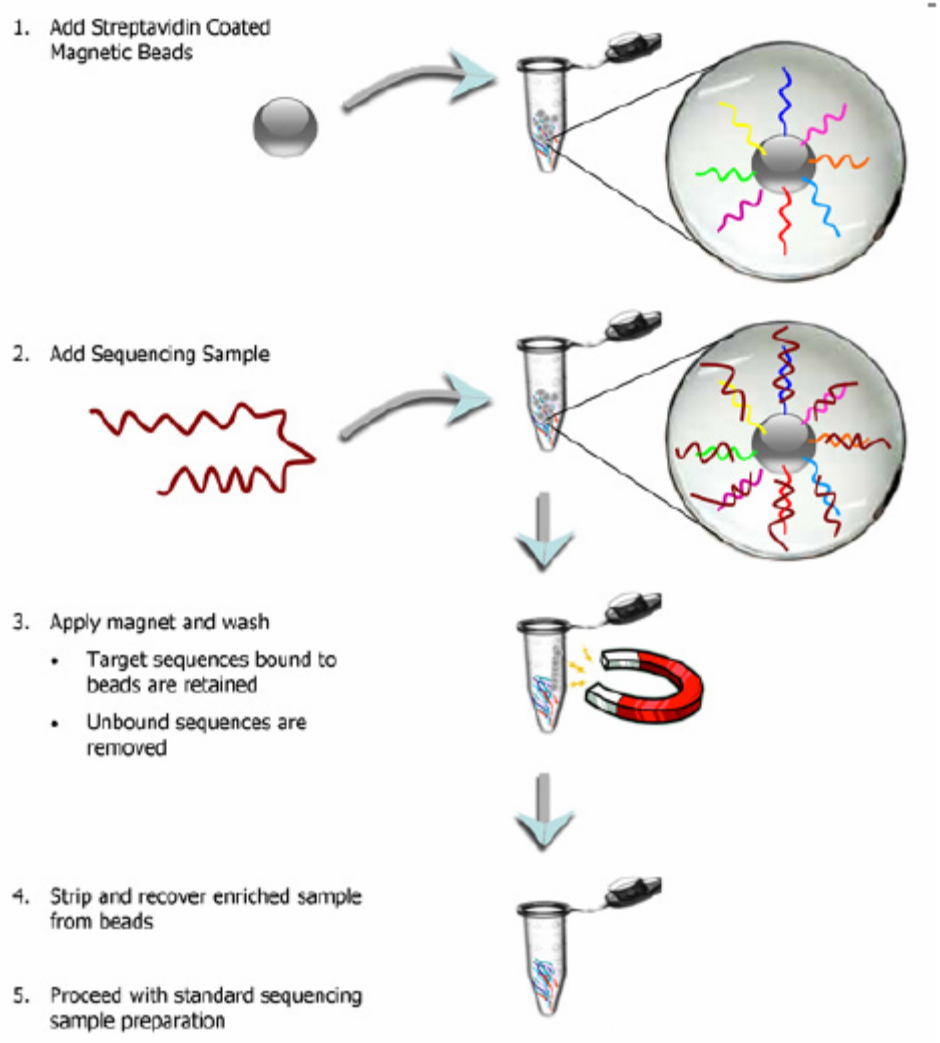


A single Nimblgen 2.1M array is capable of enrichment simultaneously more than 180,000 human coding-exon and 550 miRNA exons available in the CCDS database. The methods significantly improve the efficiency, the cost and throughput of target enrichment compared to conventional PCR-based methods.

The human exome captured library preparation consists of three major processes and lasts seven days, including library preparation for sequence capture, microarray hybridization, elution and library construction for sequencing, using Illumina Genome Analyzer can directly get high quality sequence data for the entire human exome.

This approach can find the exact genes and mutations causing several complex human diseases, such as cancer, diabetes, obesity and so on.

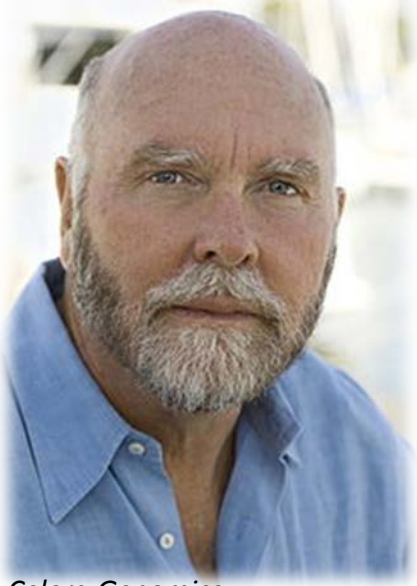
Capture Sequencing: Beads-based



Human Cancer Genes
Neonatal Genetic Disease
Human Mitochondria
Human Deafness
Human Kinome
Human X-chromosome Genes

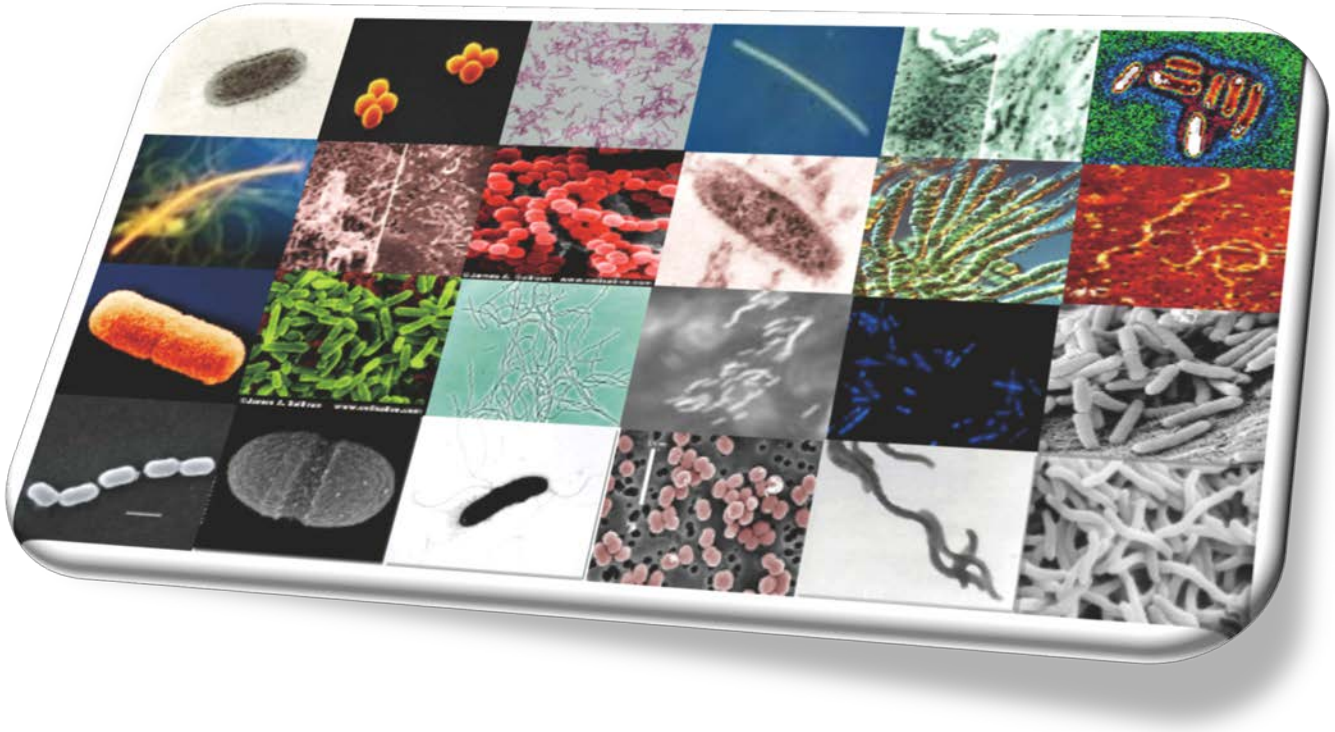
What is Metagenomics?

Craig Venter

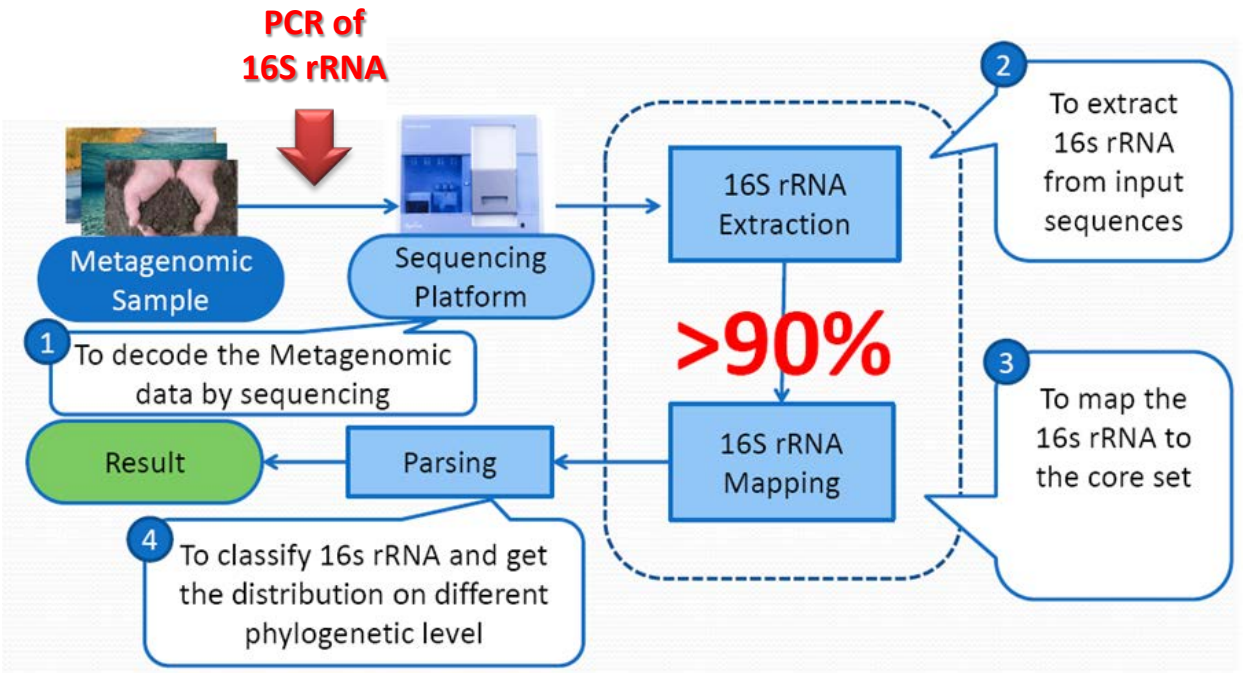


Celera Genomics
The Institute for Genomic Research
The J. Craig Venter Institute

Metagenomics is the study of metagenomes by recovering genetic materials directly from environmental or biological samples.



Classical Metagenomics Approach

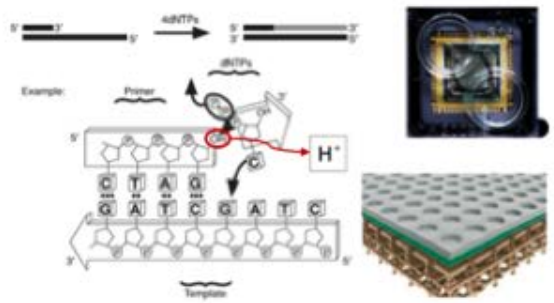


Human Microbiome Projects

Category	HMP funded	All human associated
Airways	142	212
Blood	40	117
Eye	1	7
Gastrointestinal tract	480	686
Heart	2	2
Lymph Node	1	4
Nose	3	3
Oral	345	388
Skin	292	325
Urogenital tract	286	346
Other	504	3082
All Strains	2096	5172

- 16S rRNA sequence dependent
- 1st PCR (16S rRNA family specific primer sets)
- 2nd PCR (NGS library)
- 454 Sequencing platform
- Database update

Targeted/Cancer Panel Sequencing



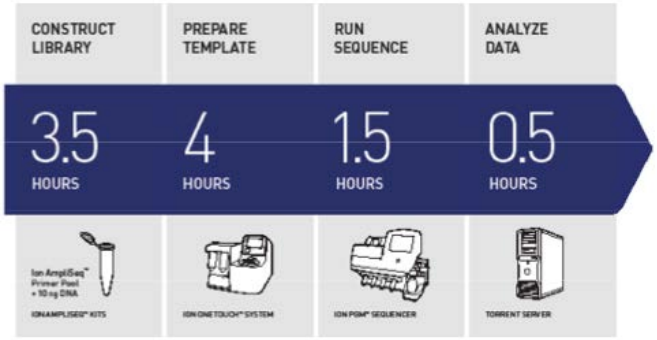






Processors: 6-core * 2
Storage: 12TB usable
Ram: 48GB

- 2.5 generation sequencing platform
- Detects the change in pH (no costly and sensitive optics)
- Lower cost and smaller instrument size
- Fast turnover rates but limited data throughput



- Benefits in clinical applications and small labs

- Ion AmpliSeq™ Ready-to-Use Panels
  - 10ng DNA
  - Cancer Hotspot Panel v1 (46 genes)
  - Cancer Hotspot Panel v2 (50 genes)
  - Comprehensive Cancer Panel (409 genes)
  - Inherited Disease Panel (325 genes, >700 i-diseases)

Ion AmpliSeq™ Cancer Panel

Ion AmpliSeq™ Cancer Panel v1

46 genes
(190 amplicons surveying 739 COSMIC mutation targets)

ABL1 (17)	EGFR (66)	HRAS (21)	NOTCH1 (8)	SMARCB1 (4)
AKT1 (1)	ERBB2 (16)	IDH1 (4)	NPM1 (3)	SMO (5)
ALK (2)	ERBB4 (11)	JAK2 (2)	NRAS (27)	SRC (1)
APC (42)	FBXW7 (7)	JAK3 (2)	PDGFRA (23)	STK11 (16)
ATM (21)	FGFR1 (1)	KDR (11)	PIK3CA (29)	TP53 (73)
BRAF (54)	FGFR2 (7)	KIT (36)	PTEN (27)	VHL (9)
CDH1 (3)	FGFR3 (11)	KRAS (34)	PTPN11 (15)	
CDKN2A (4)	FLT3 (24)	MET (9)	RB1 (9)	
CSF1R (7)	GNAS (2)	MLH1 (1)	RET (13)	
CTNNB1 (33)	HNF1A (3)	MPL (2)	SMAD4 (19)	

Amplicon length 44~124bp

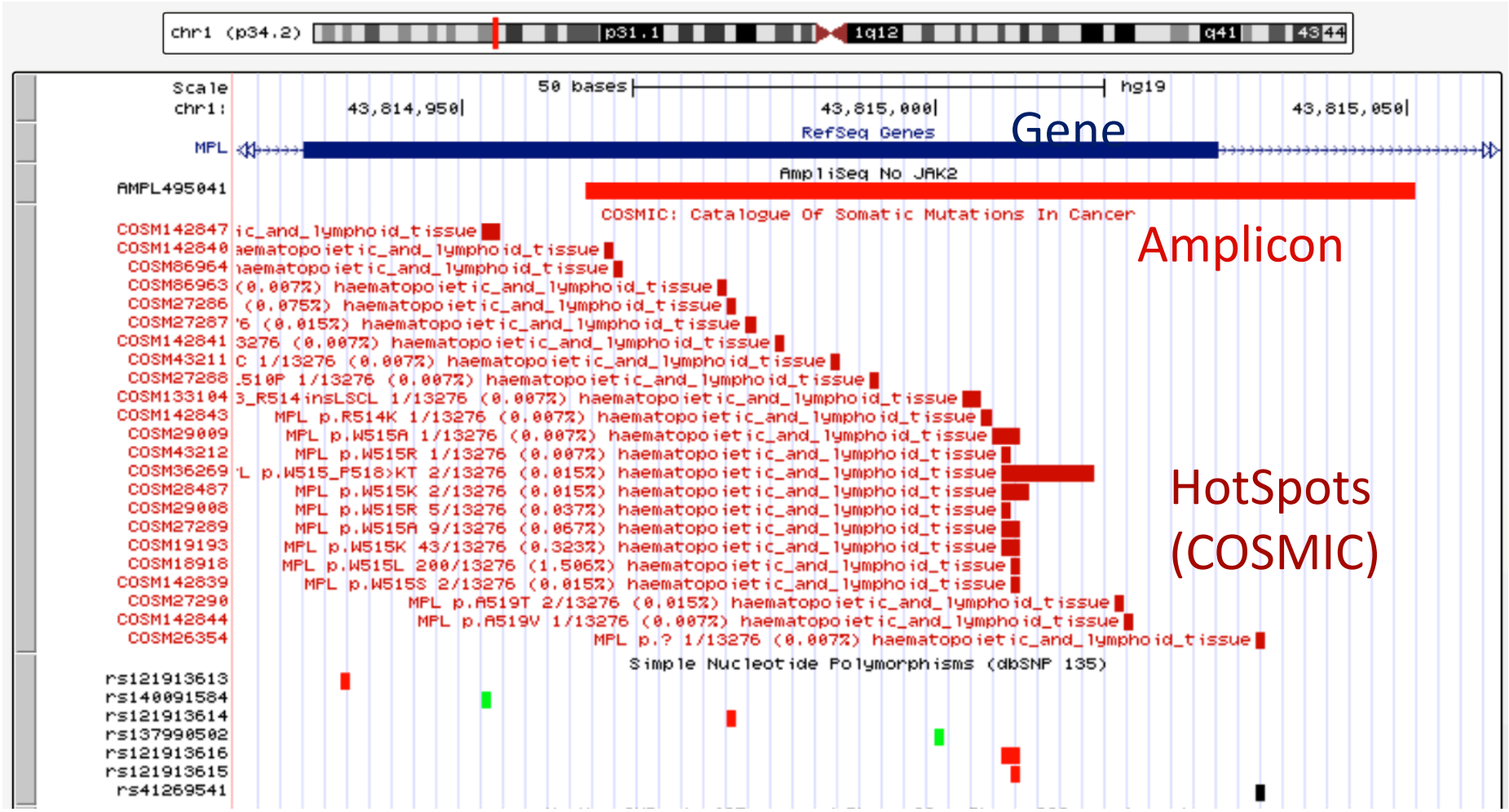
Ion AmpliSeq™ Cancer Panel v2

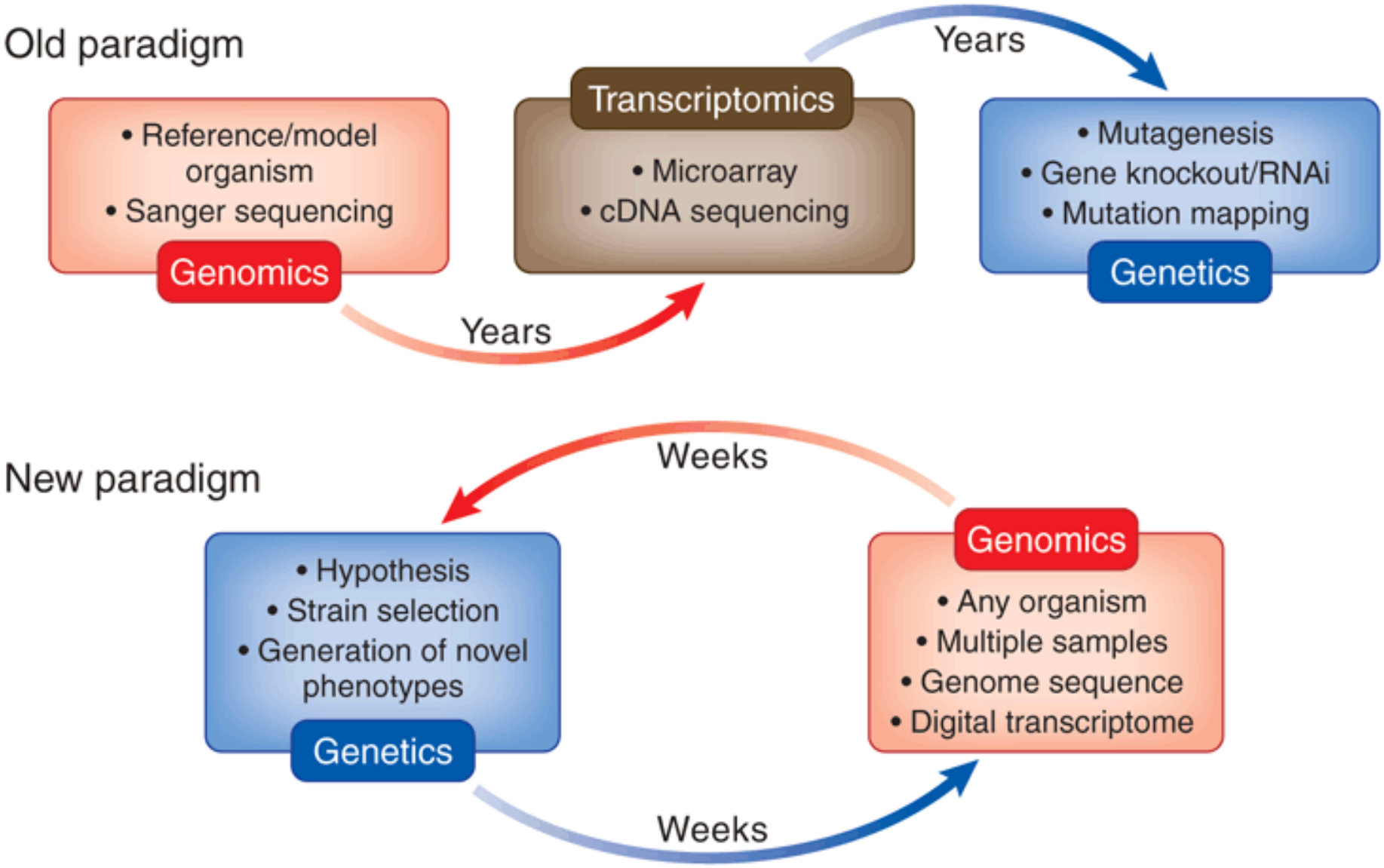
50 genes
(207 amplicons surveying 2,849 COSMIC + 6 non-COSMIC mutation targets)

ABL1 (19)	EGFR (123)	HRAS (27)	NOTCH1 (20)	SMARCB1 (11)
AKT1 (6)	ERBB2 (19)	IDH1 (15)	NPM1 (28)	SMO (5)
ALK (8)	ERBB4 (13)	IDH2 (12)	NRAS (35)	SRC (1)
APC (164)	EZH2 (11)	JAK2 (5)	PDGFRA (26)	STK11 (23)
ATM (24)	FBXW7 (25)	JAK3 (6)	PIK3CA (97)	TP53 (1150)
BRAF (77)	FGFR1 (2)	KDR (11)	PTEN (159)	VHL (124)
CDH1 (7)	FGFR2 (8)	KIT (139)	PTPN11 (28)	
CDKN2A (108)	FGFR3 (17)	KRAS (63)	RB1 (18)	
CSF1R (8)	FLT3 (30)	MET (18)	RET (17)	
CTNNB1 (73)	GNA11 (5)	MLH1 (1)	SMAD4 (31)	
	GNAQ (6)	MPL (10)		
	GNAS (12)			
	HNF1A (10)			

Amplicon length 50~141bp

Single Nucleotide Polymorphism (SNP) in Sample




Old paradigm
• Reference/model organism
• Sanger sequencing
Genomics
Years
Transcriptomics
• Microarray
• cDNA sequencing
Years
• Mutagenesis
• Gene knockout/RNAi
• Mutation mapping
Genetics
New paradigm
Weeks
• Hypothesis
• Strain selection
• Generation of novel phenotypes
Genetics
Genomics
• Any organism
• Multiple samples
• Genome sequence
• Digital transcriptome
Weeks

Mid-Term Report

Impacts of New Generation Sequencing Technologies on Biomedical Research

Guidelines:

1. You can write a report on the global impacts OR
   Just focus on the impact on any single field or application
2. Title page with your name and student number
3. Content AT LESS 4 pages (Times new Roman, size 12 font, single space)
4. Email the report to petang@mail.cgu.edu.tw before 31 Dec 2014.