Genomics/DB Course

Yeoun Jin Kim, Ph.D.
CRP-Sante

- Goals
  - To understand current genomics technologies and database structure.
  - To be able to link genomics information to proteomics research.

- Schedule
  - June 3rd afternoon session.
  - June 4th afternoon session.

- Prior Knowledge
  - Molecular biology.
  - Mass spectrometry based proteomics .

Genomics/DB Course

- Basic concept
  - Molecular biology (genome structure, cloning, sequencing)
  - Bioinformatics (data analysis)

- Human Genome Project
  - What is done and what is left
  - Technologies used in HGP

- Database landscape
  - NCBI (National Center for Biotechnology Information): www.ncbi.nlm.nih.gov
  - EBI (European Bioinformatics Institute): www.ebi.ac.uk
  - UCSC Genome bioinformatics: genome.ucsc.edu
  - UniProtKB: www.uniprot.org

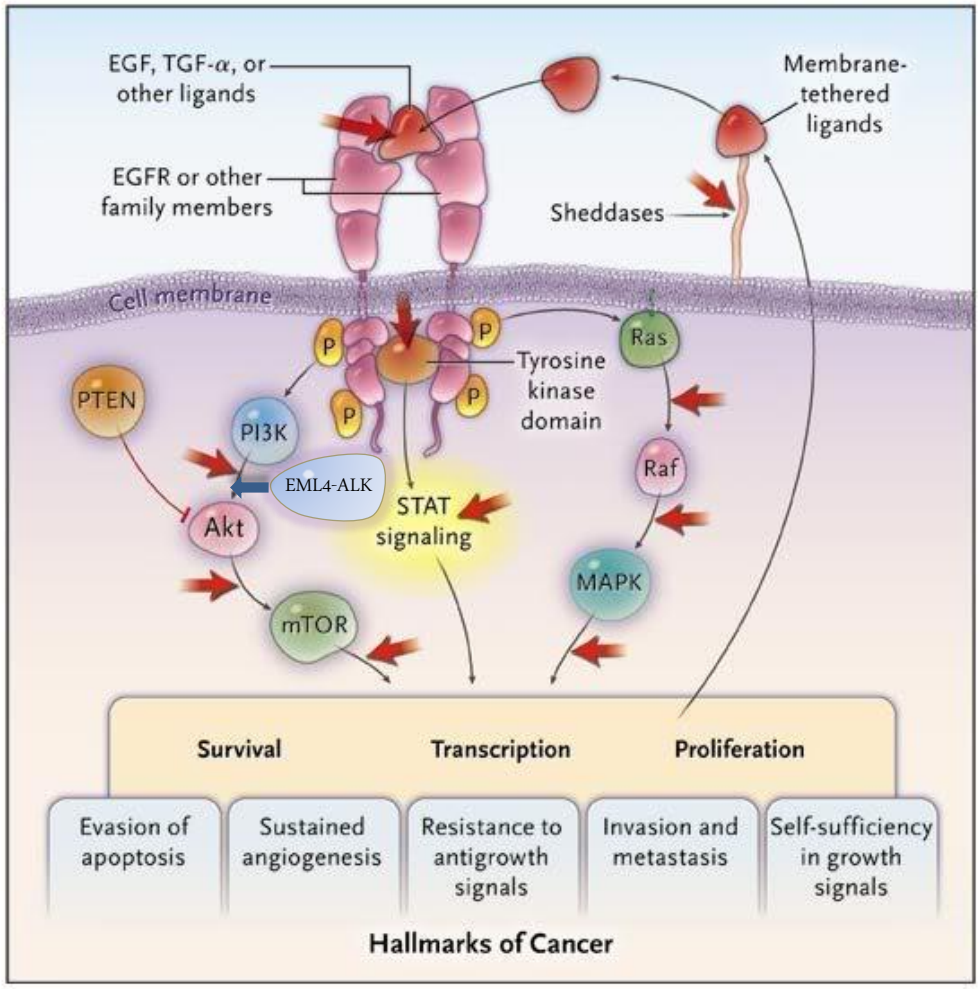
Case I

- Sequencing
  - G-I: Standard Sanger method
  - G-II: Automated sequencing method
  - G-III: Massively parallel sequencing methods
    - Pyrosequencing
    - Sequencing by synthesis
    - Sequencing by ligation
- Disease Association

Case II

Case III

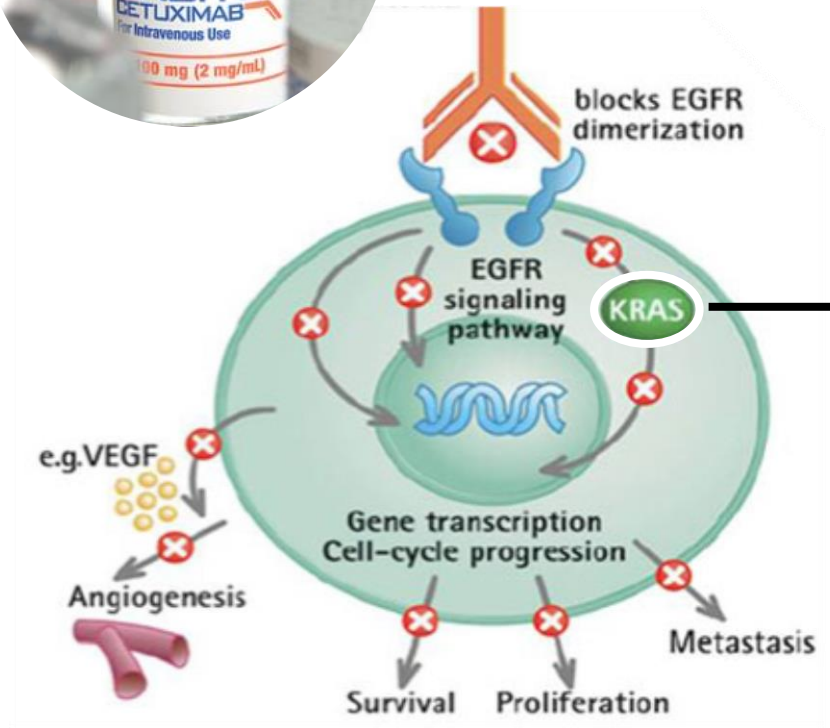
Before we start…
Why we study Genomics to Proteomics?

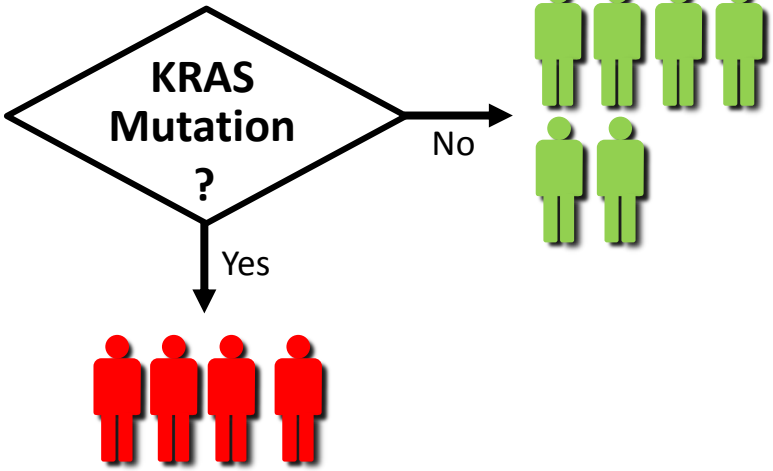


EGFR based signal pathway

Before we start…
Why we study Genomics to Proteomics?

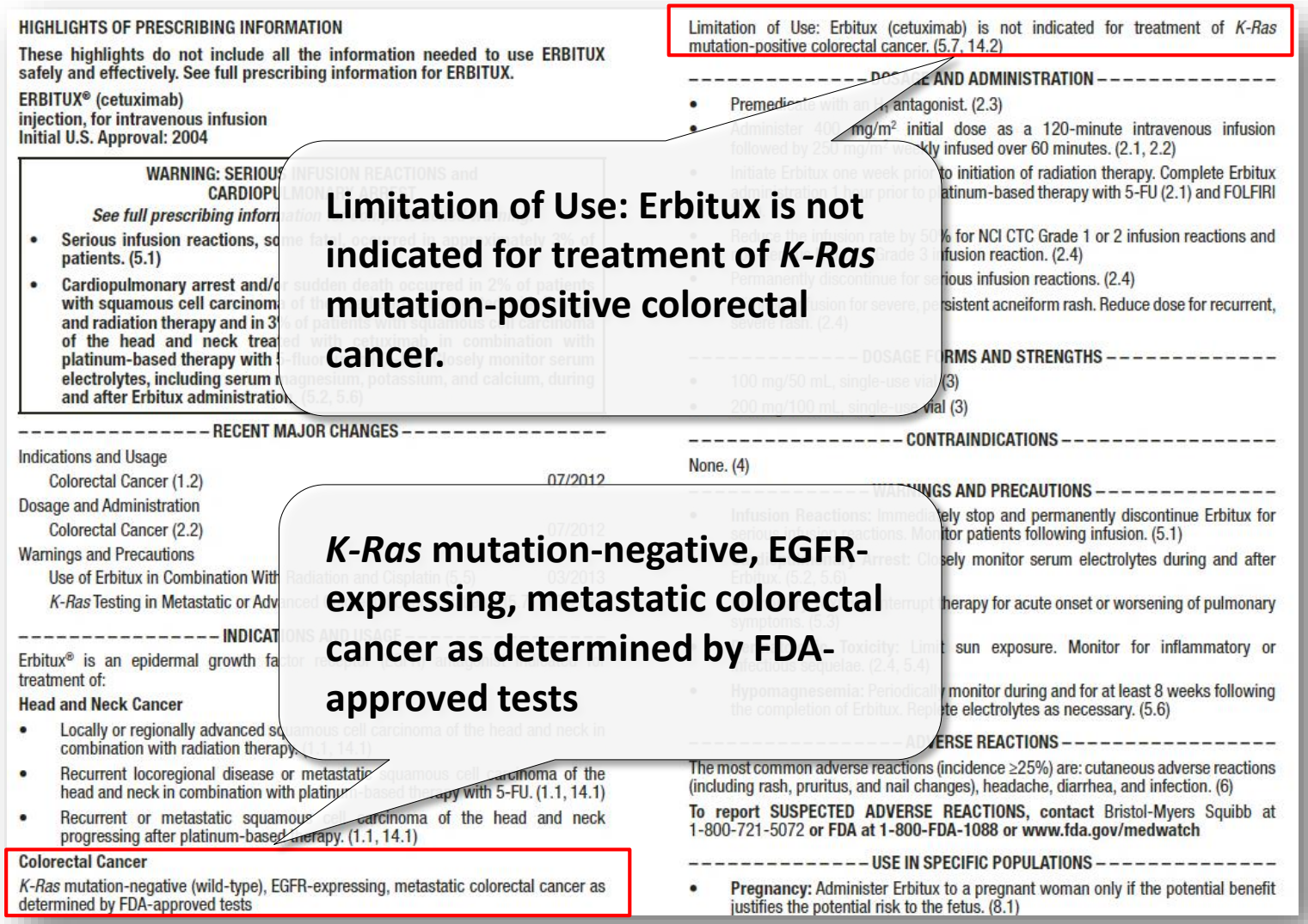
Erbitux (Cetuximab) BMS: In 2009, the FDA issued a class labeling change for Erbitux





Case of re-labeling

Erbitux (Cetuximab) BMS



HIGHLIGHTS OF PRESCRIBING INFORMATION

These highlights do not include all the information needed to use ERBITUX safely and effectively. See full prescribing information for ERBITUX.

ERBITUX® (cetuximab)
injection, for intravenous infusion
Initial U.S. Approval: 2004

WARNING: SERIOU… INFUSION REACTIONS and CARDIOPU…
See full prescribing inform…

- Serious infusion reactions, s… patients. (5.1)
- Cardiopulmonary arrest and/o… with squamous cell carcinom… and radiation therapy and in 3… of the head and neck trea… platinum-based therapy with 5-… electrolytes, including serum m…, magnesium, potassium, and calcium, during and after Erbitux administratio… (5.2, 5.6)

RECENT MAJOR CHANGES

Indications and Usage
Colorectal Cancer (1.2) 07/2012
Dosage and Administration
Colorectal Cancer (2.2) 07/2012
Warnings and Precautions
Use of Erbitux in Combination With… Radiation and Cisplatin (5.5) 03/2013
K-Ras Testing in Metastatic or Adv…

INDICAT… AND USAGE

Erbitux® is an epidermal growth fa… treatment of:

Head and Neck Cancer

- Locally or regionally advanced s… combination with radiation therapy. (1.1, 14.1)
- Recurrent locoregional disease or metastatic … squamous cell carcinoma of the head and neck in combination with platinu… -based therapy with 5-FU. (1.1, 14.1)
- Recurrent or metastatic squamous … carcinoma of the head and neck progressing after platinum-based therapy. (1.1, 14.1)

Colorectal Cancer

K-Ras mutation-negative (wild-type), EGFR-expressing, metastatic colorectal cancer as determined by FDA-approved tests

Limitation of Use: Erbitux (cetuximab) is not indicated for treatment of *K-Ras* mutation-positive colorectal cancer. (5.7, 14.2)

DOSAGE AND ADMINISTRATION

- Premedi… with an H1 antagonist. (2.3)
- Administer 400 mg/m² initial dose as a 120-minute intravenous infusion followed by 250 mg/m² weekly infused over 60 minutes. (2.1, 2.2)
- Initiate Erbitux one week prior to initiation of radiation therapy. Complete Erbitux administration 1 hour prior to platinum-based therapy with 5-FU (2.1) and FOLFIRI
- Reduce infusion rate by 50% for NCI CTC Grade 1 or 2 infusion reactions and Grade 3 infusion reaction. (2.4)
- Permanently discontinue for serious infusion reactions. (2.4)
- Withhold for severe, persistent acneiform rash. Reduce dose for recurrent, … (2.4)

DOSAGE FORMS AND STRENGTHS

- 100 mg/50 mL, single-use vial (3)
- 200 mg/100 mL, single-use vial (3)

CONTRAINDICATIONS

None. (4)

WARNINGS AND PRECAUTIONS

- Infusion Reactions: Immediately stop and permanently discontinue Erbitux for severe infusion reactions. Monitor patients following infusion. (5.1)
- Cardiopulmonary Arrest: Closely monitor serum electrolytes during and after Erbitux. (5.2, 5.6)
- Pulmonary Toxicity: Interrupt therapy for acute onset or worsening of pulmonary symptoms. (5.3)
- Dermatologic Toxicity: Limit sun exposure. Monitor for inflammatory or infectious sequelae. (2.4, 5.4)
- Hypomagnesemia: Periodically monitor during and for at least 8 weeks following the completion of Erbitux. Replete electrolytes as necessary. (5.6)

ADVERSE REACTIONS

The most common adverse reactions (incidence ≥25%) are: cutaneous adverse reactions (including rash, pruritus, and nail changes), headache, diarrhea, and infection. (6)

To report SUSPECTED ADVERSE REACTIONS, contact Bristol-Myers Squibb at 1-800-721-5072 **or FDA at 1-800-FDA-1088 or www.fda.gov/medwatch**

USE IN SPECIFIC POPULATIONS

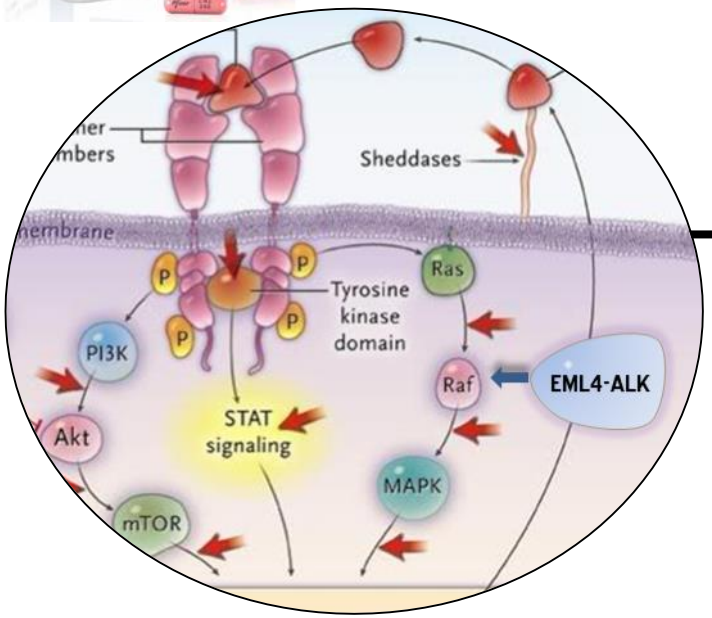
- **Pregnancy:** Administer Erbitux to a pregnant woman only if the potential benefit justifies the potential risk to the fetus. (8.1)

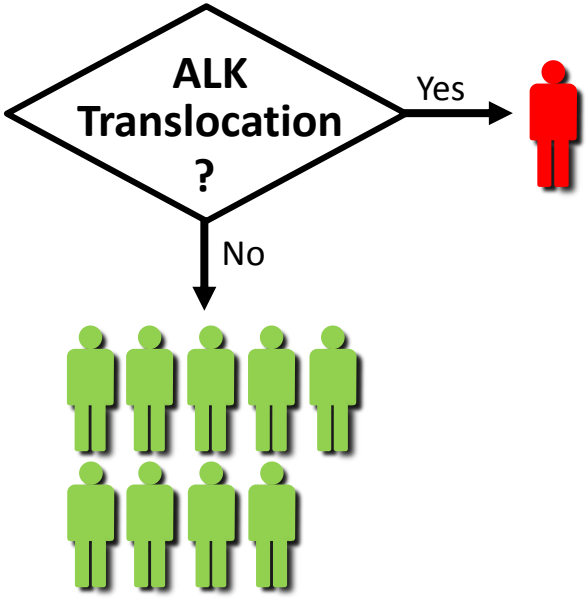
Limitation of Use: Erbitux is not indicated for treatment of *K-Ras* mutation-positive colorectal cancer.

***K-Ras* mutation-negative, EGFR-expressing, metastatic colorectal cancer as determined by FDA-approved tests**

Xalkori (Crizotinib) Pfizer

In 2011, the FDA 2011, granted accelerated approval to Xalkori with an concurrent approval of ALK Break-Apart FISH Probe Kit (Abbott).





R.S. Herbst *et al.*, *NEJM* 2008,

Xalkori (Crizotinib) Pfizer

withhold XALKORI until recovery to less than or equal to Grade 1, then resume XALKORI at 250 mg once daily. Permanently discontinue XALKORI if Grade 3 QTc prolongation recurs *[see Dosage and Administration (2.2) and Clinical Pharmacology (12.4)]*.

5.4 ALK Testing
Detection of ALK-positive NSCLC using an FDA-approved test, indicated for this use, is necessary for selection of patients for treatment with XALKORI *[see Clinical Studies (14)]*.

Assessment for ALK-positive NSCLC should be performed by laboratories with demonstrated proficiency in the specific technology being utilized. Improper assay performance can lead to unreliable test results.

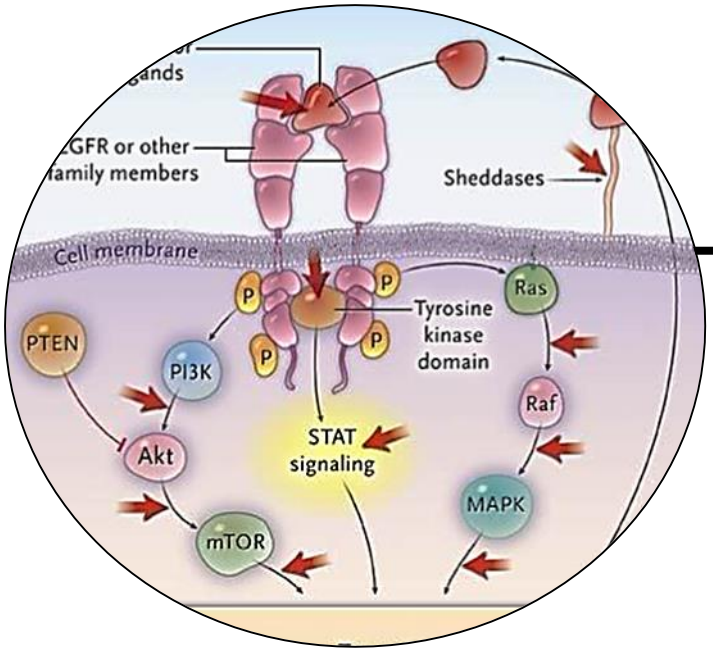
Refer to an FDA-approved test's package insert for instructions on the identification of patients eligible for treatment with XALKORI.

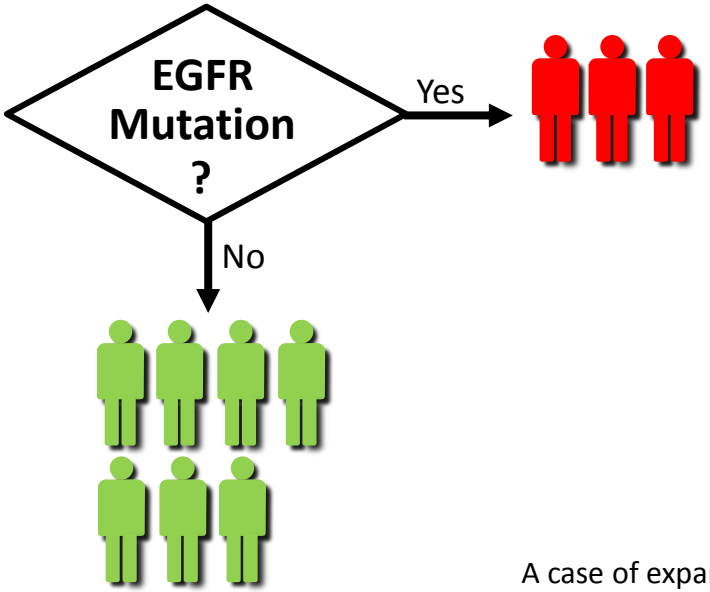
5.5 Pregnancy
XALKORI can cause fetal harm when administered to a pregnant woman based on its mechanism of action. In nonclinical studies in rats, crizotinib was embryotoxic and fetotoxic at exposures similar to and above those observed in humans at the recommended clinical dose of 250 mg twice daily. There are no adequate and well-controlled studies in pregnant women using XALKORI. If this drug is used during pregnancy, or if the patient becomes pregnant while taking this drug, apprise the patient of the potential hazard to a fetus *[see Use in Specific Populations (8.1)]*.

6. ADVERSE REACTIONS
Because clinical trials are conducted under widely varying conditions, adverse reaction rates observed in the clinical trials of a drug cannot be directly compared to rates in the clinical trials of another drug and may not reflect the rates observed in practice.

Tarceva (Erlotinib) Roche

In 2013, the FDA approved Tarceva for the first-line treatment of metastatic NSCLC with EFGR mutation test (Cobas).





A case of expanding market

HIGHLIGHTS OF PRESCRIBING INFORMATION

These highlights do not include all the information needed to use ERBITUX safely and effectively. See full prescribing information for ERBITUX.

ERBITUX® (cetuximab)
injection, for intravenous infusion
Initial U.S. Approval: 2004

WARNING: SERIOUS INFUSION REACTIONS and CARDIOPULMONARY ARREST

See full prescribing information for complete boxed warning.

- **Serious infusion reactions, some fatal, occurred in approximately 3% of patients. (5.1)**
- **Cardiopulmonary arrest and/or sudden death occurred in 2% of patients with squamous cell carcinoma of the head and neck treated with Erbitux and radiation therapy and in 3% of patients with squamous cell carcinoma of the head and neck treated with cetuximab in combination with platinum-based therapy with 5-fluorouracil (5-FU). Closely monitor serum electrolytes, including serum magnesium, potassium, and calcium, during and after Erbitux administration. (5.2, 5.6)**

RECENT MAJOR CHANGES

Indications and Usage	
Colorectal Cancer (1.2)	07/2012
Dosage and Administration	
Colorectal Cancer (2.2)	07/2012
Warnings and Precautions	
Use of Erbitux in Combination With Radiation and Cisplatin (5.5)	03/2013
K-Ras Testing in Metastatic or Advanced Colorectal Cancer Patients (5.7)	07/2012

INDICATIONS AND USAGE

Erbitux® is an epidermal growth factor receptor (EGFR) antagonist indicated for treatment of:

Head and Neck Cancer

- Locally or regionally advanced squamous cell carcinoma of the head and neck in combination with radiation therapy. (1.1, 14.1)
- Recurrent locoregional disease or metastatic squamous cell carcinoma of the head and neck in combination with platinum-based therapy with 5-FU. (1.1, 14.1)
- Recurrent or metastatic squamous cell carcinoma of the head and neck progressing after platinum-based therapy. (1.1, 14.1)

Colorectal Cancer

K-Ras mutation-negative (wild-type), EGFR-expressing, metastatic colorectal cancer as determined by FDA-approved tests

- in combination with FOLFIRI for first-line treatment,
- in combination with irinotecan in patients who are refractory to irinotecan-based chemotherapy,
- as a single agent in patients who have failed oxaliplatin- and irinotecan-based chemotherapy or who are intolerant to irinotecan. (1.2, 5.7, 12.1, 14.2)

Limitation of Use: Erbitux (cetuximab) is not indicated for treatment of *K-Ras* mutation-positive colorectal cancer. (5.7, 14.2)

DOSAGE AND ADMINISTRATION

- Premedicate with an H_1 antagonist. (2.3)
- Administer 400 mg/m^2 initial dose as a 120-minute intravenous infusion followed by 250 mg/m^2 weekly infused over 60 minutes. (2.1, 2.2)
- Initiate Erbitux one week prior to initiation of radiation therapy. Complete Erbitux administration 1 hour prior to platinum-based therapy with 5-FU (2.1) and FOLFIRI (2.2).
- Reduce the infusion rate by 50% for NCI CTC Grade 1 or 2 infusion reactions and non-serious NCI CTC Grade 3 infusion reaction. (2.4)
- Permanently discontinue for serious infusion reactions. (2.4)
- Withhold infusion for severe, persistent acneiform rash. Reduce dose for recurrent, severe rash. (2.4)

DOSAGE FORMS AND STRENGTHS

- 100 mg/50 mL, single-use vial (3)
- 200 mg/100 mL, single-use vial (3)

CONTRAINDICATIONS

None. (4)

WARNINGS AND PRECAUTIONS

- **Infusion Reactions:** Immediately stop and permanently discontinue Erbitux for serious infusion reactions. Monitor patients following infusion. (5.1)
- **Cardiopulmonary Arrest:** Closely monitor serum electrolytes during and after Erbitux. (5.2, 5.6)
- **Pulmonary Toxicity:** Interrupt therapy for acute onset or worsening of pulmonary symptoms. (5.3)
- **Dermatologic Toxicity:** Limit sun exposure. Monitor for inflammatory or infectious sequelae. (2.4, 5.4)
- **Hypomagnesemia:** Periodically monitor during and for at least 8 weeks following the completion of Erbitux. Replete electrolytes as necessary. (5.6)

ADVERSE REACTIONS

The most common adverse reactions (incidence ≥25%) are: cutaneous adverse reactions (including rash, pruritus, and nail changes), headache, diarrhea, and infection. (6)

To report SUSPECTED ADVERSE REACTIONS, contact Bristol-Myers Squibb at 1-800-721-5072 **or FDA at 1-800-FDA-1088 or www.fda.gov/medwatch**

USE IN SPECIFIC POPULATIONS

- **Pregnancy:** Administer Erbitux to a pregnant woman only if the potential benefit justifies the potential risk to the fetus. (8.1)
- **Nursing Mothers:** Discontinue nursing during and for 60 days following treatment with Erbitux. (8.3)

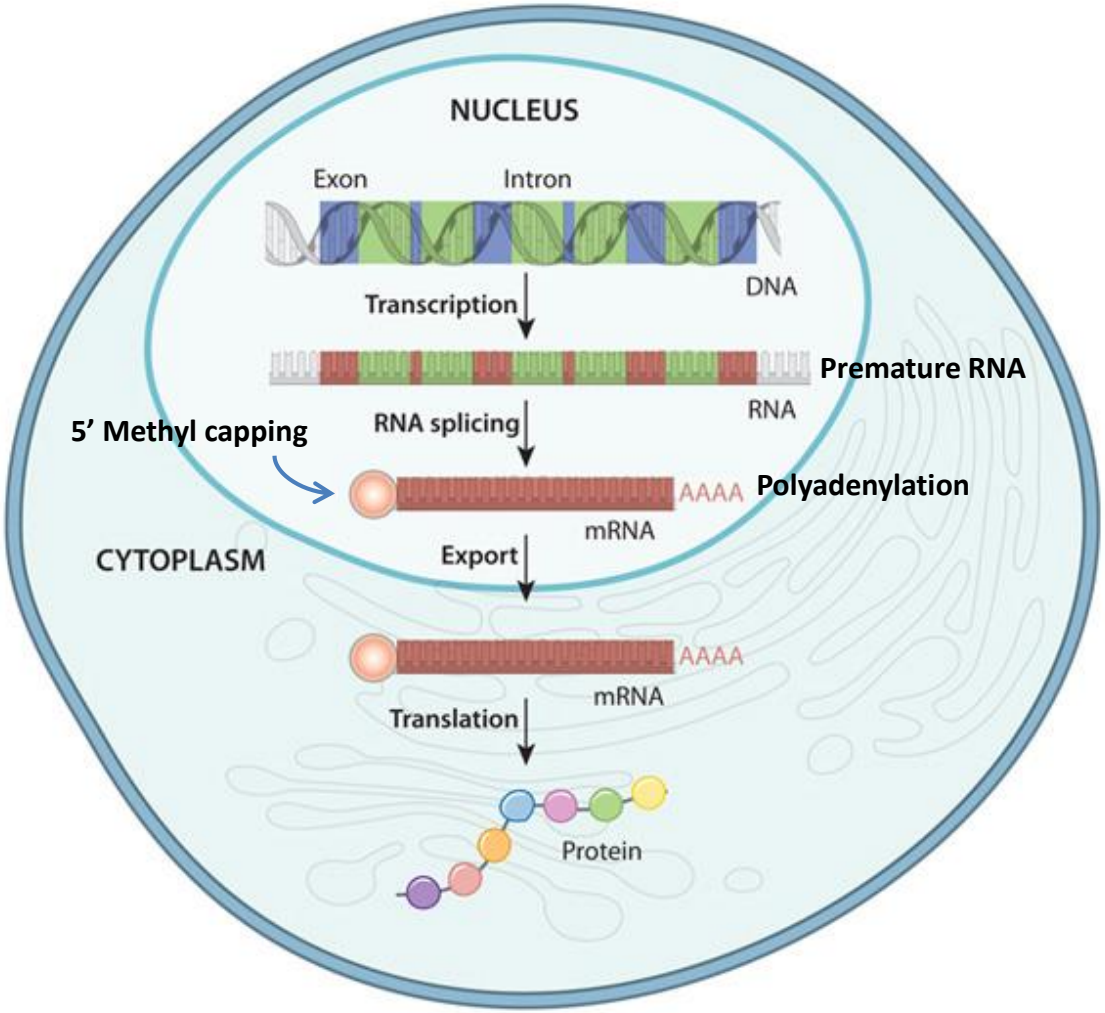
See 17 for PATIENT COUNSELING INFORMATION

Revised: 03/2013

BRCA1/2 mutations and Breast Cancer

In 2013…

Linking the genetic components to the functional consequences in proteome level and the impact on the clinical decision point-personalized medicine.



Genome

Transcriptome

Epigenome

Proteome

Brief history of Human Genome Project

• Project goals: To identify all genes in human DNA, determine the sequences of the 3 billion chemical base pairs that make up human DNA, store this information in databases, improve tools for data analysis

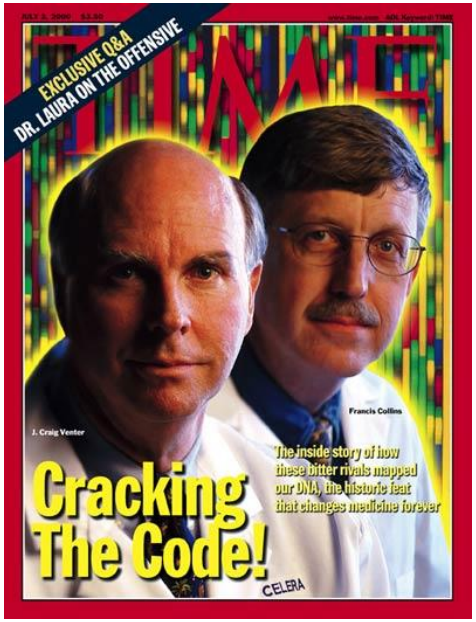
• Estimated budget: $ 3 billion
• Starting date: 1990
• Projected completion date for draft genome sequence: 2005

• Public venture coordinated by the U.S. Department of Energy and the National Institutes of Health and led by Francis Collins with major partners in UK, France, Japan, Germany, China …..

• Private venture led by Craig Venter at Celera.

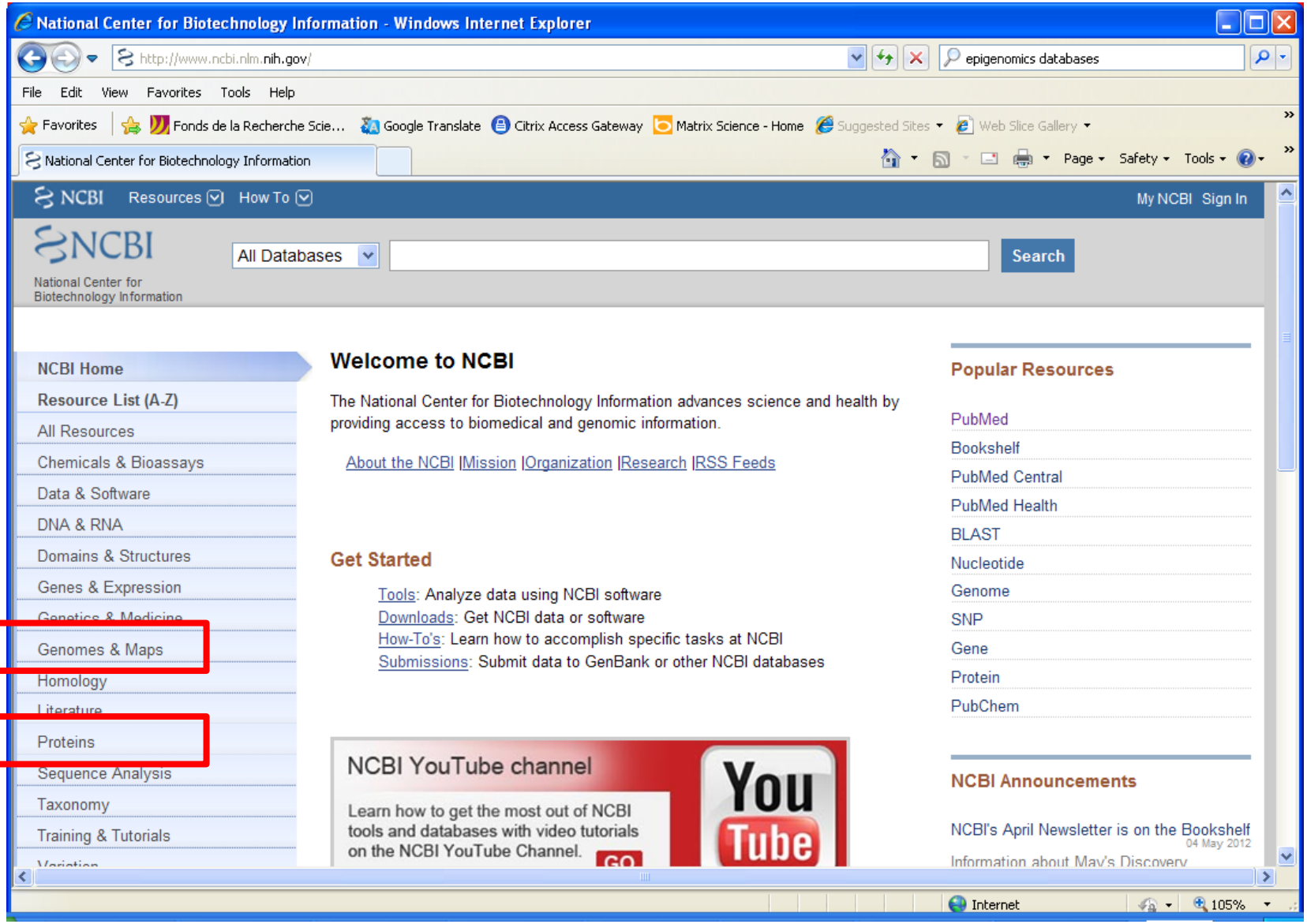
• Draft sequence announced jointly May, 2000 at the White House.

• Draft sequences published in Nature (public) and Science (Celera) in Feb 2001, four years ahead of schedule.

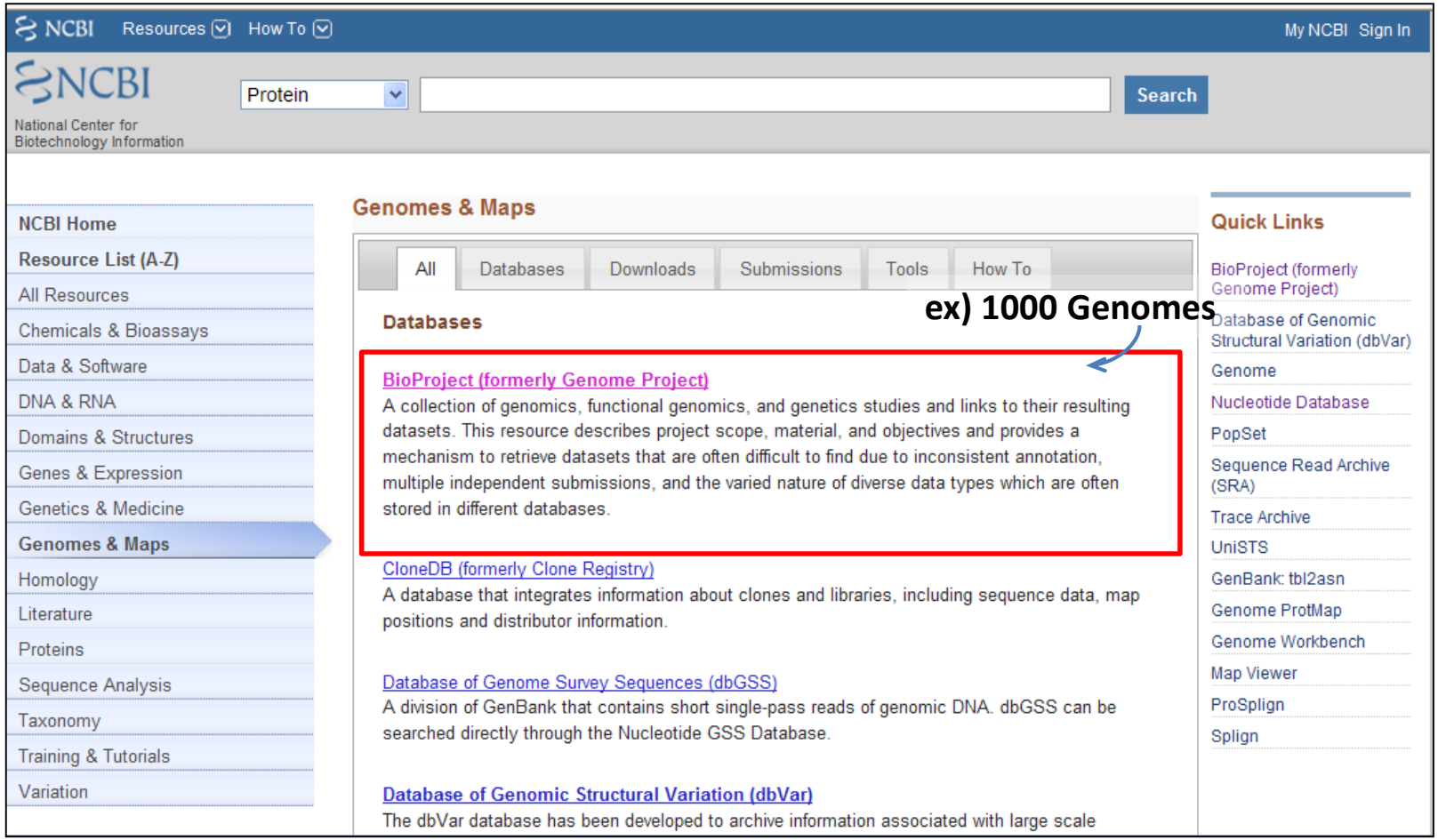


As a result!

http://www.ncbi.nlm.nih.gov/



As a result!


NCBI
Resources
How To
My NCBI
Sign In
NCBI
National Center for Biotechnology Information
Protein
Search
NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation
Genomes & Maps
All
Databases
Downloads
Submissions
Tools
How To
Databases
ex) 1000 Genomes
BioProject (formerly Genome Project)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.
CloneDB (formerly Clone Registry)
A database that integrates information about clones and libraries, including sequence data, map positions and distributor information.
Database of Genome Survey Sequences (dbGSS)
A division of GenBank that contains short single-pass reads of genomic DNA. dbGSS can be searched directly through the Nucleotide GSS Database.
Database of Genomic Structural Variation (dbVar)
The dbVar database has been developed to archive information associated with large scale
Quick Links
BioProject (formerly Genome Project)
Database of Genomic Structural Variation (dbVar)
Genome
Nucleotide Database
PopSet
Sequence Read Archive (SRA)
Trace Archive
UniSTS
GenBank: tbl2asn
Genome ProtMap
Genome Workbench
Map Viewer
ProSplign
Splign

As a result!

1

Genome/Human Genome
-Gene Database/RefSeq
-dbSNP
rs# (reference snp id)

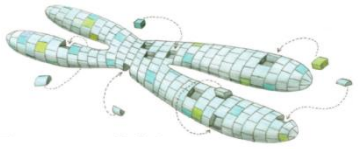
2

The Wellcome Trust Sanger Institute

The Genome Institute at Washington University

The European Bioinformatics Institute

The National Center for Biotechnology Information



"The goal of the GRC is to correct misassembled regions, to close remaining gaps, and to provide alternate assemblies of structurally variant positions (loci) in the genome."

Needs of Reference Genome?

Genomes & Maps - Site Guide - NCBI

Genome
Contains sequence and map data from the whole genomes of over 1000 organisms. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life (bacteria, archaea, and eukaryota) are represented, as well as many viruses, phages, viroids, plasmids, and organelles.

Genome Reference Consortium (GRC)
The Genome Reference Consortium (GRC) maintains responsibility for the human and mouse reference genomes. Members consist of The Genome Center at Washington University, the Wellcome Trust Sanger Institute, the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI). The GRC works to correct misrepresented loci and to close remaining assembly gaps. In addition, the GRC seeks to provide alternate assemblies for complex or structurally variant genomic loci. At the GRC website (http://www.genomereference.org), the public can view genomic regions currently under review, report genome-related problems and contact the GRC.

HIV-1, Human Protein Interaction Database
A database of known interactions of HIV-1 proteins with proteins from human hosts. It provides annotated bibliographies of published reports of protein interactions, with links to the corresponding PubMed records and sequence data.

Influenza Virus
A compilation of data from the NIAID Influenza Genome Sequencing Project and GenBank. It provides tools for flu sequence analysis, annotation and submission to GenBank. This resource also has links to other flu sequence resources, and publications and general information about flu viruses.

3

Nucleotide Database
A collection of nucleotide sequences from several sources, including GenBank, RefSeq, the Third Party Annotation (TPA) database, and PDB. Searching the Nucleotide Database will yield available results from each of its component databases.

Reference: Baseline

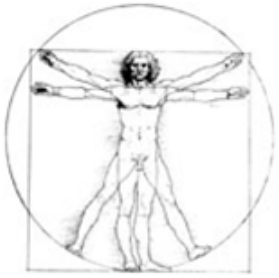
GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Human | Mouse | Zebrafish | Paper Supplemental Data

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

Genome Assemblies

The GRC has built tools to facilitate the curation of genome assemblies based on the sequence overlaps of long, high quality sequences (Clones and PCR products, not short sequence reads). The GRC currently supports production of assemblies for human, mouse or zebrafish. If your assembly data fits this model and you are interested in using these tools please contact us using the 'Contact Us' page.



Human

The human genome assembly was produced as part of the Human Genome Project (HGP). The previous assembly (NCBI36) was the last one produced by the HGP and was described in 2004 (PMID: 15496913); this was the starting point for the GRC. The assembly is based largely on assembling overlapping clone sequences.

Human assembly information

Current Major Assembly	GRCh37
Regions with Alternate Loci	3
Assembly N50	46,395,641 bp
Remaining Gaps	357
Patch Release version	p8
Patches Released	Fix: 69; Novel: 71

Needs of Reference Genome?

1) Scientific communication/reference
2) Alignment of the sub-genome/genes

Reference: Baseline

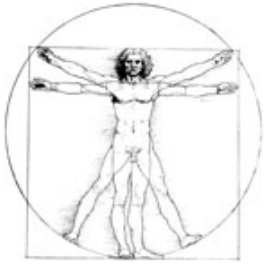
GRC Home | Data | Help | Report an Issue | Contact Us | Credits | Curators Only

Human | Mouse | Zebrafish | Paper Supplemental Data

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

Genome Assemblies

The GRC has built tools to facilitate the curation of genome assemblies based on the sequence overlaps of long, high quality sequences (Clones and PCR products, not short sequence reads). The GRC currently supports production of assemblies for human, mouse or zebrafish. If your assembly data fits this model and you are interested in using these tools please contact us using the 'Contact Us' page. Subscribe to the grc-announce email list to receive email notification for all GRC assembly updates.



Human

The human genome assembly was produced as part of the Human Genome Project (HGP). The previous assembly (NCBI36) was the last one produced by the HGP and was described in 2004 (PMID: 15496913); this was the starting point for the GRC. The assembly is based largely on assembling overlapping clone sequences.

Human assembly information

Current Major Assembly	GRCh38
Regions with Alternate Loci	178
Assembly N50	67,794,873 bp
Remaining Gaps	875

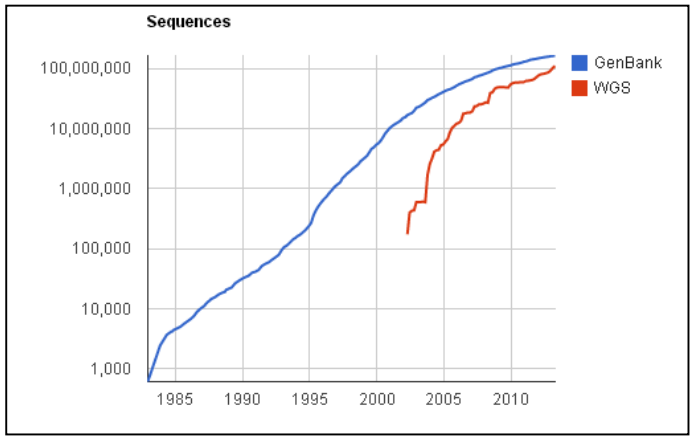
Visit 1000 genome webpage
FAQ site
http://www.1000genomes.org/faq

Q: “Which reference assembly do you use?”

GenBank

GenBank: The NIH genetic sequence database, an annotated collection of all publicly available DNA sequences*. There are 164 136 731 sequence records in the Genbank as of April 2013.
GenBank is part of the International Nucleotide Sequence Database Collaboration (INSDC).

GenBank at NCBI
DNA DataBank of Japan (DDBJ)
European Molecular Biology Laboratory (EMBL)



GI number ("gi") : a series of digits that are assigned consecutively to each sequence record processed by NCBI.

VERSION: the accession number of the database record followed by a dot and a version number.

The GI number has been used for many years by NCBI to track sequence histories in GenBank and the other sequence databases it maintains. The VERSION system of identifiers was adopted in February 1999 by the INSDC. The two systems of identifiers run in parallel to each other.

When any change is made to a sequence, it receives a new GI number AND an increase to its version number.

*Guidance and Introduction:
http://nar.oxfordjournals.org/content/41/D1/D36.full.pdf+html

RefSeq : annotated information of DNA, RNA, and protein. constructed from sequence data submitted to the INSDC.

" Reference Sequence Project"

Annotation for RefSeq

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Intermediate genomic assemblies of BAC and/or WGS sequence data
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NS_	Genomic	Environmental sequence
NZ_[b]	Genomic	Unfinished WGS**
NM_	mRNA	Transcript products; mature messenger RNA (mRNA) transcripts.
NR_	RNA	Non-coding transcripts including structural RNAs, transcribed pseudogenes, and others.
XM_[c]	mRNA	Predicted* model
XR_[c]	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_	Protein	Protein products; no corresponding transcript record provided. Primarily used for bacterial, viral, and mitochondrial records.
XP_	Protein	Predicted model, associated with an XM_ accession
ZP_	Protein	Predicted model, annotated on NZ_ genomic records

**PREDICTED: Not curated.* Automatically provided based on GenBank sequence data; limited or partial support for the transcript or protein. A portion of the transcript or protein may reflect an *ab initio* annotation prediction that was submitted to GenBank.

***WGS: Not curated.* The RefSeq record represents a collection of whole genome shotgun (WGS) sequences. This status code is applied to genomic records
.

Example

Search: “homo sapiens SAA1” in nucleotide section.

RefSeq (13)

Transcript variant 1, mRNA

Check RefSeq accession, GI, version

Search the previous version

CDS for protein

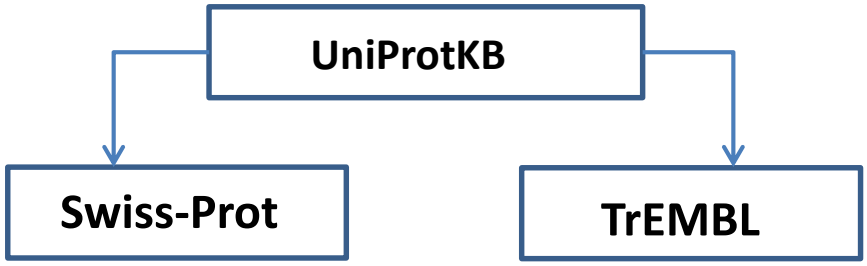
NP_accession

Summary

Entry	mRNA	Protein	mRNA version	GI	GENE	GENE
TR variant 1, mRNA	NM_000331	NP_000322.2	NM_000331.4	GI:295821191	GeneID:6288	HGNC:10513
TR variant 2, mRNA	NM_199161	NP_954630.1	NM_199161.3	GI:295821190	GeneID:6288	HGNC:10513
TR variant 3, mRNA	NM_001178006	NP_001171477.1	NM_001178006.1	GI:295821192	GeneID:6288	HGNC:10513

Protein DB

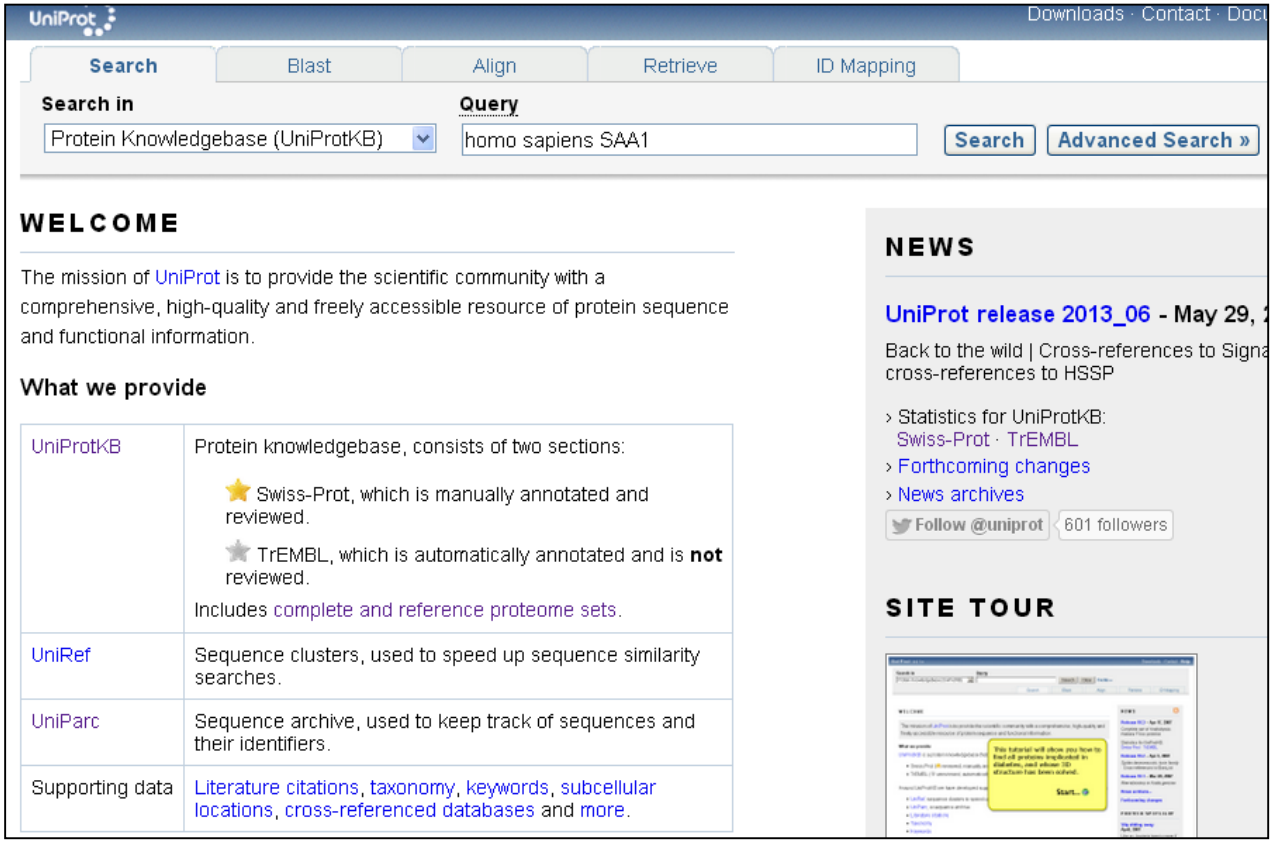
protein DB with evidence in protein level

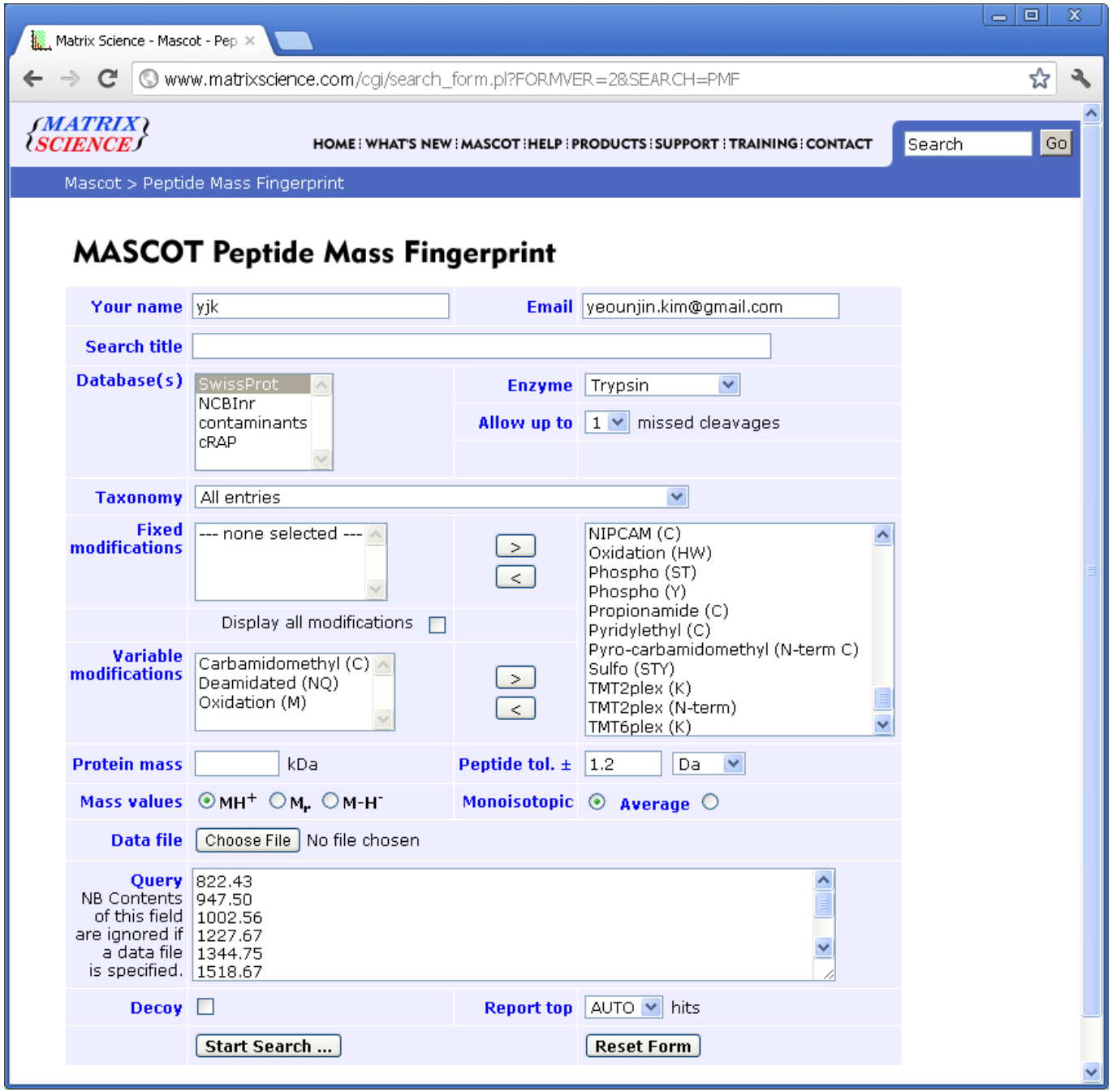


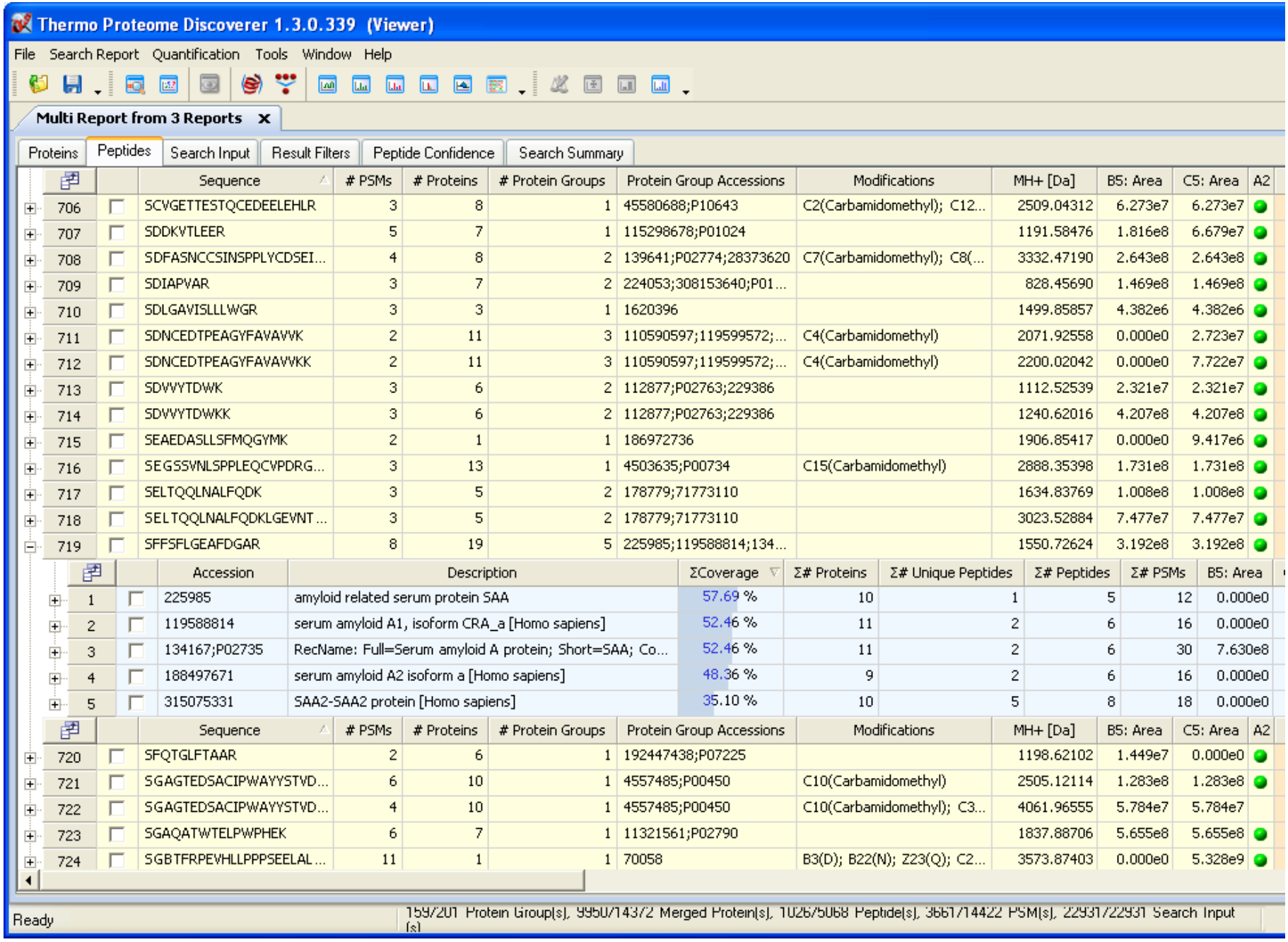
www.uniprot.org

Human Curation

Translated EMBL

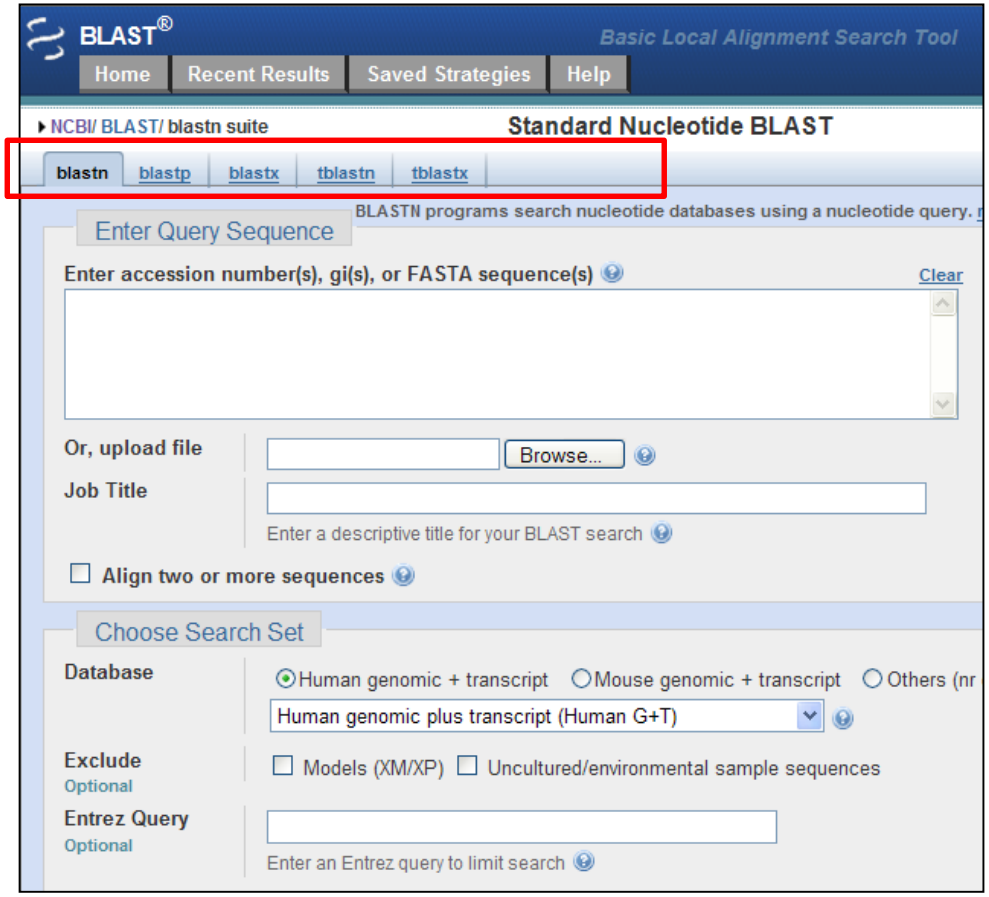



Matrix Science - Mascot - Pep
www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=PMF
MATRIX SCIENCE
HOME | WHAT'S NEW | MASCOT | HELP | PRODUCTS | SUPPORT | TRAINING | CONTACT
Search
Go
Mascot > Peptide Mass Fingerprint
MASCOT Peptide Mass Fingerprint
Your name
yjk
Email
yeounjin.kim@gmail.com
Search title
Database(s)
SwissProt
NCBInr
contaminants
cRAP
Enzyme
Trypsin
Allow up to
1
missed cleavages
Taxonomy
All entries
Fixed modifications
--- none selected ---
Display all modifications
Variable modifications
Carbamidomethyl (C)
Deamidated (NQ)
Oxidation (M)
NIPCAM (C)
Oxidation (HW)
Phospho (ST)
Phospho (Y)
Propionamide (C)
Pyridylethyl (C)
Pyro-carbamidomethyl (N-term C)
Sulfo (STY)
TMT2plex (K)
TMT2plex (N-term)
TMT6plex (K)
Protein mass
kDa
Peptide tol. ±
1.2
Da
Mass values
MH+
Mr
M-H-
Monoisotopic
Average
Data file
Choose File
No file chosen
Query
NB Contents of this field are ignored if a data file is specified.
822.43
947.50
1002.56
1227.67
1344.75
1518.67
Decoy
Report top
AUTO
hits
Start Search ...
Reset Form


Thermo Proteome Discoverer 1.3.0.339 (Viewer)
File Search Report Quantification Tools Window Help
Multi Report from 3 Reports
Proteins Peptides Search Input Result Filters Peptide Confidence Search Summary
Sequence # PSMs # Proteins # Protein Groups Protein Group Accessions Modifications MH+ [Da] B5: Area C5: Area A2
706 SCVGETTESTQCEDEELEHLR 3 8 1 45580688;P10643 C2(Carbamidomethyl); C12... 2509.04312 6.273e7 6.273e7
707 SDDKVTLEER 5 7 1 115298678;P01024 1191.58476 1.816e8 6.679e7
708 SDFASNCCSINSPPLYCDSEI... 4 8 2 139641;P02774;28373620 C7(Carbamidomethyl); C8(... 3332.47190 2.643e8 2.643e8
709 SDIAPVAR 3 7 2 224053;308153640;P01... 828.45690 1.469e8 1.469e8
710 SDLGAVISLLLWGR 3 3 1 1620396 1499.85857 4.382e6 4.382e6
711 SDNCEDTPEAGYFAVAVVK 2 11 3 110590597;119599572;... C4(Carbamidomethyl) 2071.92558 0.000e0 2.723e7
712 SDNCEDTPEAGYFAVAVVKK 2 11 3 110590597;119599572;... C4(Carbamidomethyl) 2200.02042 0.000e0 7.722e7
713 SDVVYTDWK 3 6 2 112877;P02763;229386 1112.52539 2.321e7 2.321e7
714 SDVVYTDWKK 3 6 2 112877;P02763;229386 1240.62016 4.207e8 4.207e8
715 SEAEDASLLSFMQGYMK 2 1 1 186972736 1906.85417 0.000e0 9.417e6
716 SEGSSVNLSPPLEQCVPDRG... 3 13 1 4503635;P00734 C15(Carbamidomethyl) 2888.35398 1.731e8 1.731e8
717 SELTQQLNALFQDK 3 5 2 178779;71773110 1634.83769 1.008e8 1.008e8
718 SELTQQLNALFQDKLGEVNT... 3 5 2 178779;71773110 3023.52884 7.477e7 7.477e7
719 SFFSFLGEAFDGAR 8 19 5 225985;119588814;134... 1550.72624 3.192e8 3.192e8
Accession Description ΣCoverage Σ# Proteins Σ# Unique Peptides Σ# Peptides Σ# PSMs B5: Area
1 225985 amyloid related serum protein SAA 57.69 % 10 1 5 12 0.000e0
2 119588814 serum amyloid A1, isoform CRA_a [Homo sapiens] 52.46 % 11 2 6 16 0.000e0
3 134167;P02735 RecName: Full=Serum amyloid A protein; Short=SAA; Co... 52.46 % 11 2 6 30 7.630e8
4 188497671 serum amyloid A2 isoform a [Homo sapiens] 48.36 % 9 2 6 16 0.000e0
5 315075331 SAA2-SAA2 protein [Homo sapiens] 35.10 % 10 5 8 18 0.000e0
Sequence # PSMs # Proteins # Protein Groups Protein Group Accessions Modifications MH+ [Da] B5: Area C5: Area A2
720 SFQTGLFTAAR 2 6 1 192447438;P07225 1198.62102 1.449e7 0.000e0
721 SGAGTEDSACIPWAYYSTVD... 6 10 1 4557485;P00450 C10(Carbamidomethyl) 2505.12114 1.283e8 1.283e8
722 SGAGTEDSACIPWAYYSTVD... 4 10 1 4557485;P00450 C10(Carbamidomethyl); C3... 4061.96555 5.784e7 5.784e7
723 SGAQATWTELPWPHEK 6 7 1 11321561;P02790 1837.88706 5.655e8 5.655e8
724 SGBTFRPEVHLLPPPSEELAL... 11 1 1 70058 B3(D); B22(N); Z23(Q); C2... 3573.87403 0.000e0 5.328e9
Ready
159/201 Protein Group(s), 9950/14372 Merged Protein(s), 1026/5068 Peptide(s), 3661/14422 PSM(s), 22931/22931 Search Input (s)

BLAST (The Basic Local Alignment Search Tool)

BLAST finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

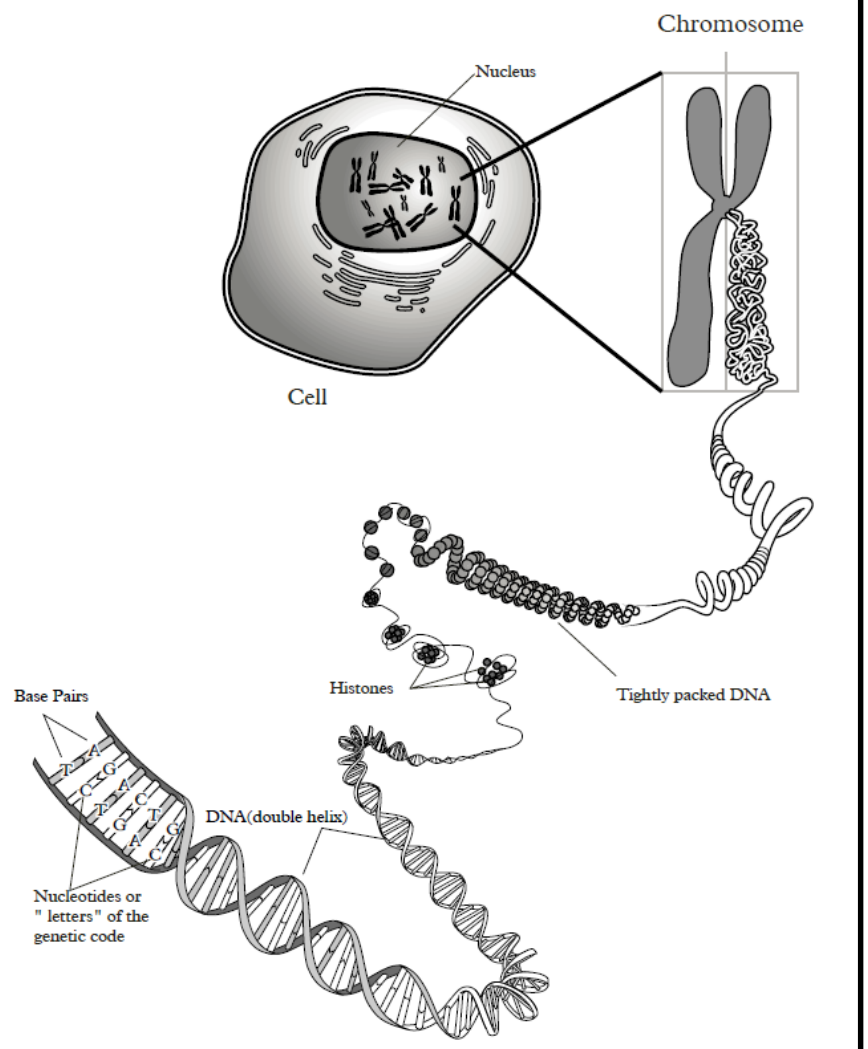


blastn: nucleotide-nucleotide
blastp: protein-protein
blastx: nucleotide 6-frame translation-protein (both strands)
tblastn: Protein-nucleotide 6-frame translation
tblastx: nucleotide 6-frame translation-nucleotide 6-frame translation

TCT AAC GCT AGT TGA AA
T CTA ACG CTA GTT GAA A
TC TAA CGC TAG TTG AAA

AAAGTTGATCGCAATCT

Human Genome Structure



Source: National Human Genome Research Institute

-3,000,000,000 base pairs (bp)

-24 chromosomes: 1-22, X, Y

-21,500 ~ 24,000 genes

-Coding DNA: only 2% of the genome encodes genes

-SNPs (single nucleotide polymorphisms)

-Mutations (change of amino acids)

-INDELs (insertion or deletions of 1~10Kb)

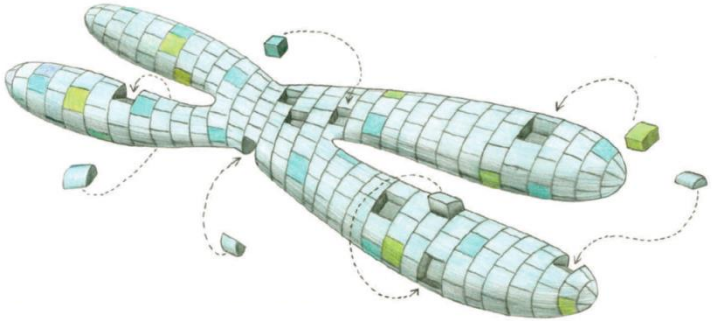
-Structural variation (rearrangement or deletion of >10Kb)

-Inversions (an entire fragment swaps ends)

-Amplification (multiple repeated copies of a genome segment)

-Human genome sequenced in the large scale project serves scientists as a “reference” to which we can compare other human genomes…

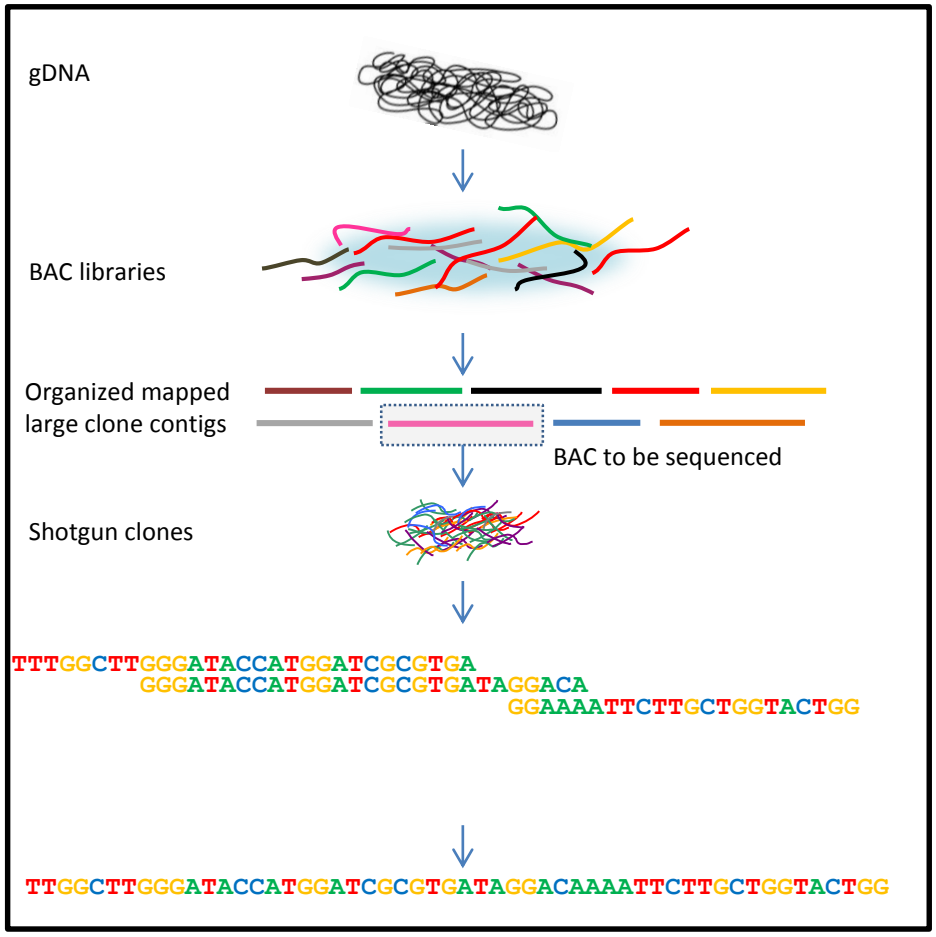
Still on going



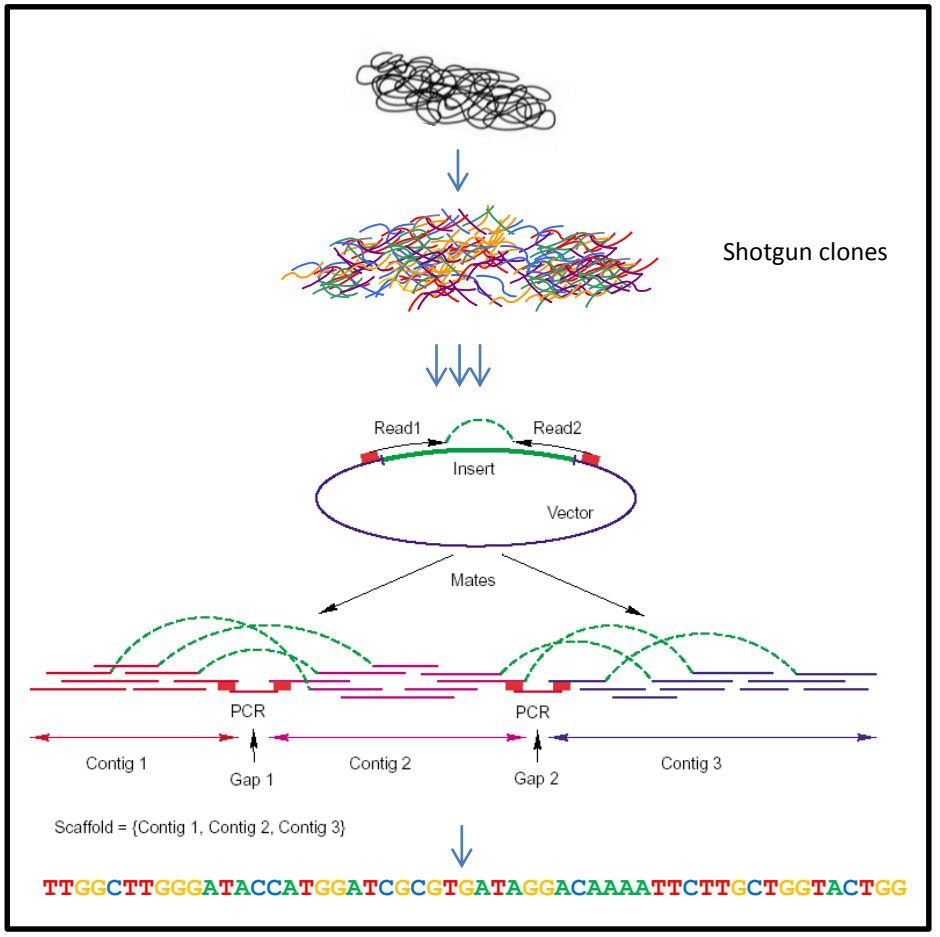
- Though the HGP is finished, analyses of the data will continue for many years.
- Re-sequencing: In general, human genomes are ~99.99% identical. However, we have little knowledge about the level of "normal" human variation beyond SNPs (amplification/deletion and inter-/intra-chromosomal rearrangements, etc.).
- Having a reference genome enables us to investigate variation in additional human genomes. A lot of current sequencing effort is spent on re-sequencing genomes of known species
  - Individual humans (1000 Genomes Project)
  - Experimental organisms – looking for genetic variation, copy number variation
- Challenge: to align millions of sequence reads to a reference genome with some % of mismatches to accurately call SNPs , indels, repeated sequences
- Disease related somatic mutations
  - TCGA consortium
  - Personalized medicine and companion diagnostics
- http://genome.ucsc.edu/
  - Exercise: Search for TMPRSS11E in GRCh36 vs GRCh37 **CASE I**

Sequencing strategies used in HGP

Public effort



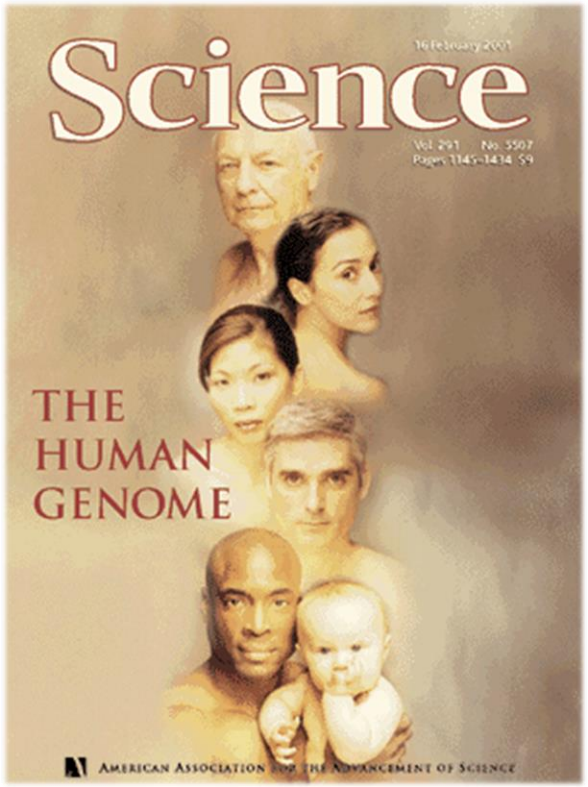
Celera



Sequencing strategies used in HGP

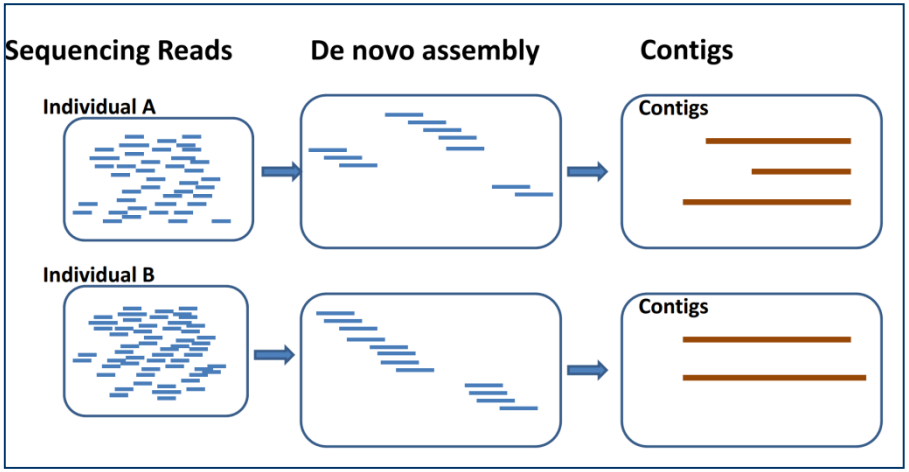
Public effort

Celera

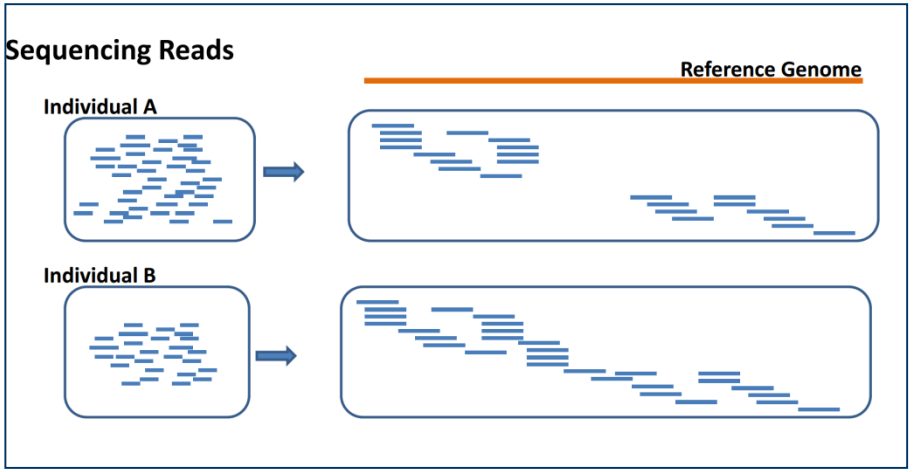


Assembly vs. Alignment

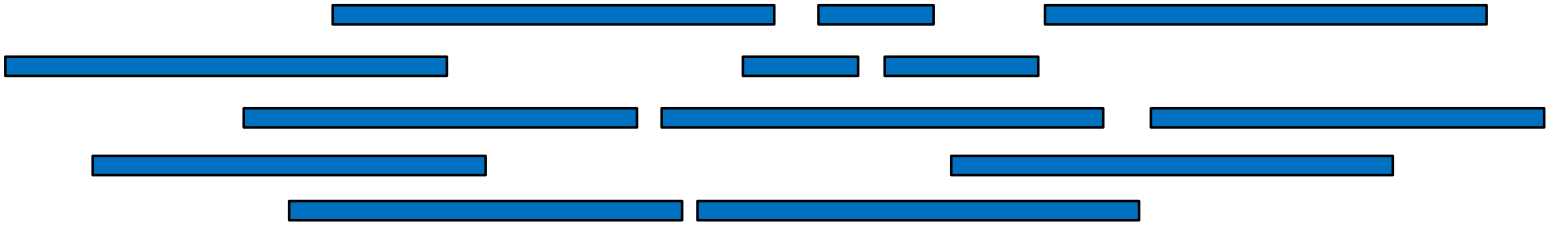
Assembly

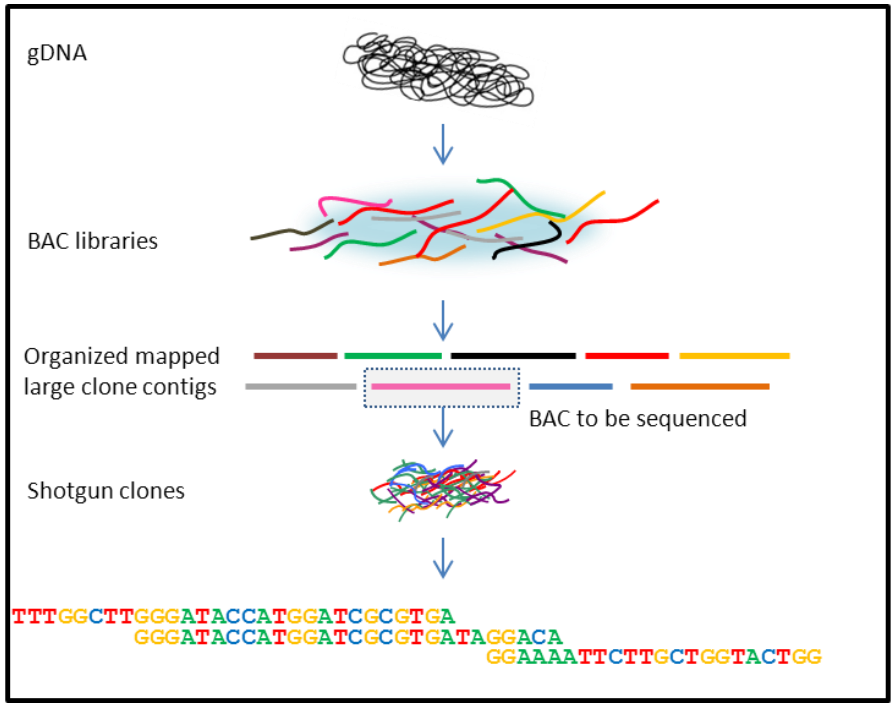


Alignment

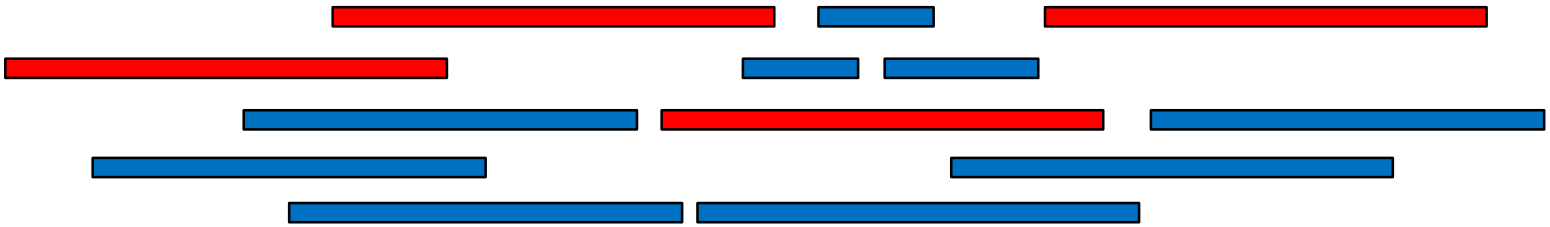


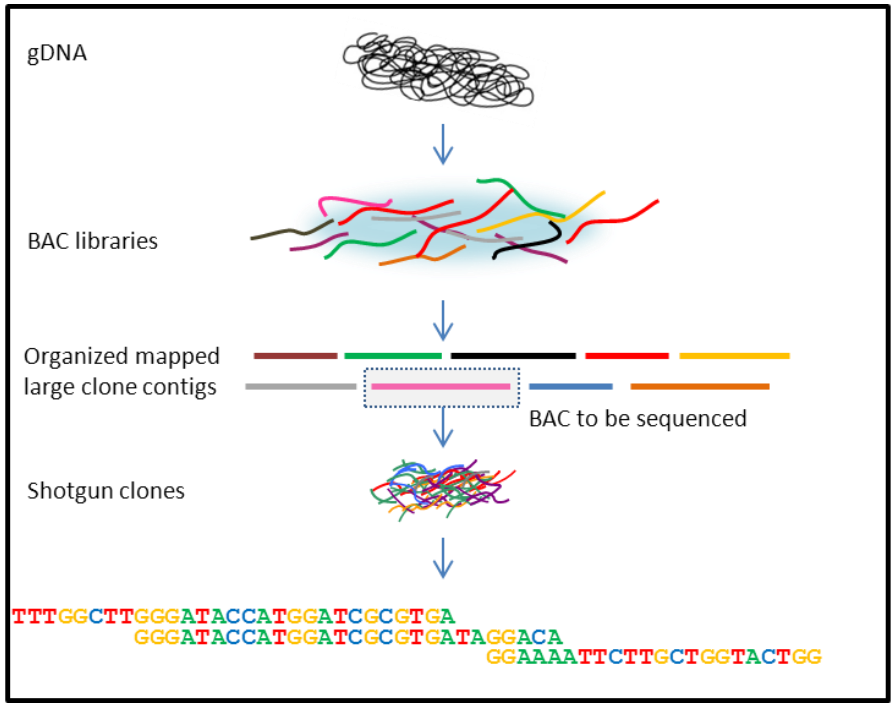
Minimal Tiling Path



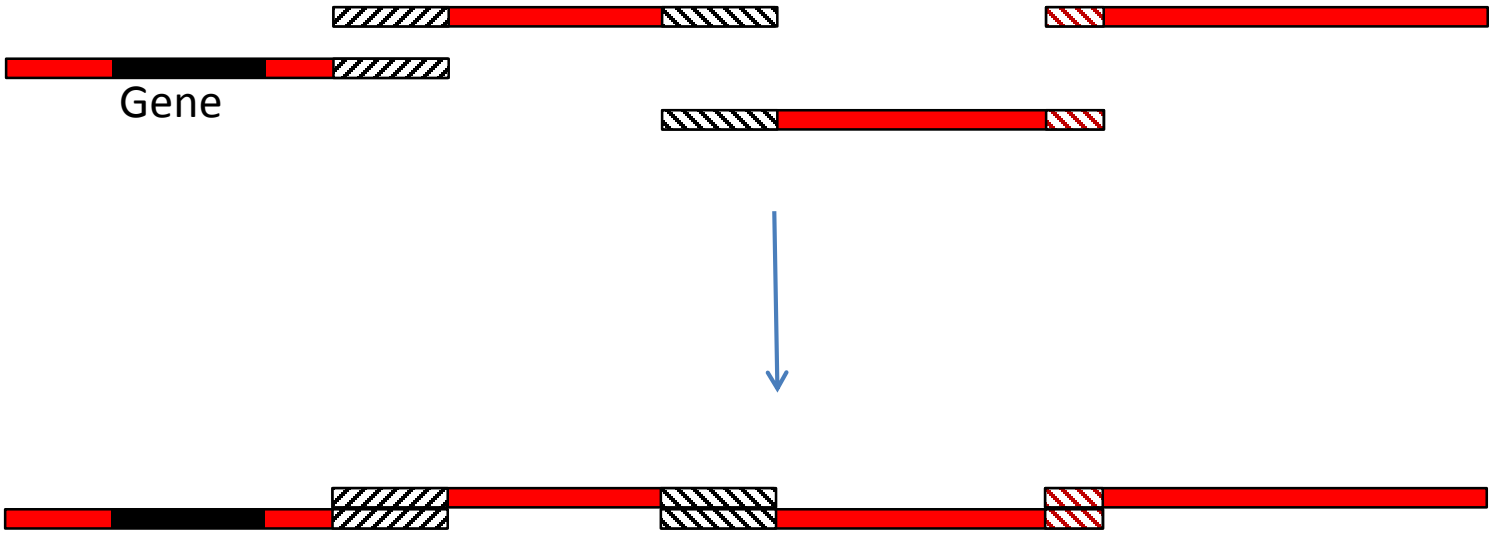


Minimal Tiling Path

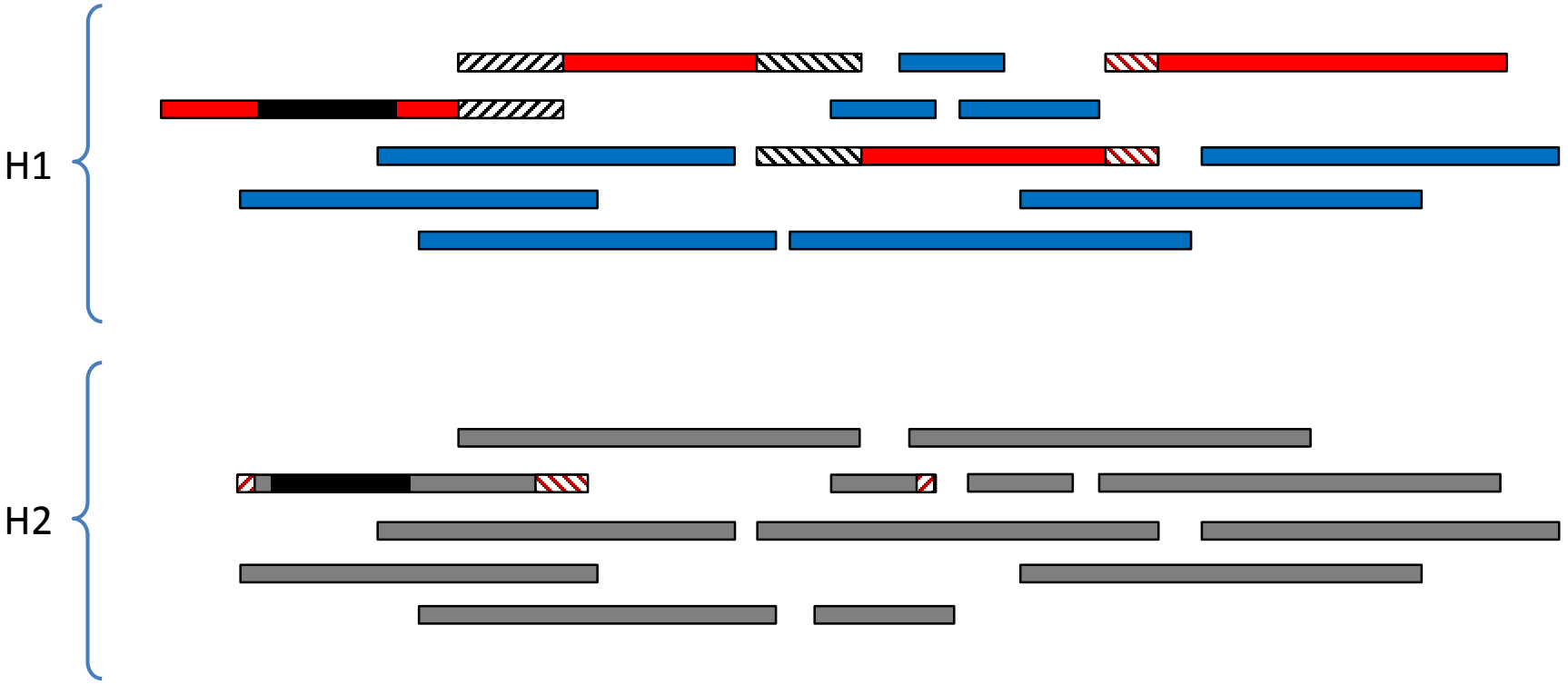




Assembly



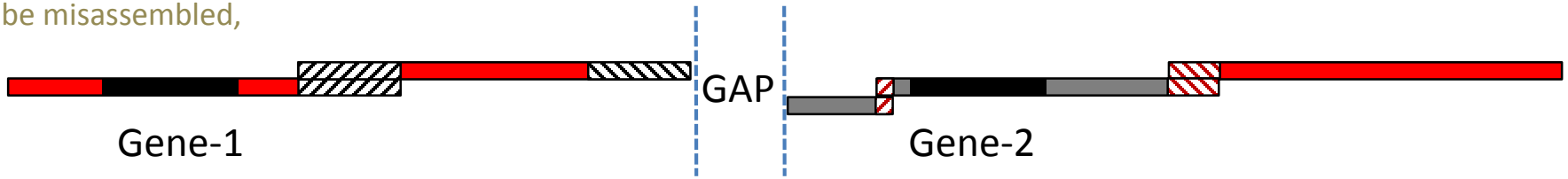
Misassembly



Misassembly

Instead of this,

Can be misassembled,



Misassembled genome in this case results in gene duplication and gap.

CASE I

Closing the GAP

Home Genomes Blat Tables Gene Sorter PCR Session FAQ Help

Human (*Homo sapiens*) Genome Browser Gateway

The UCSC Genome Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Software Copyright (c) The Regents of the University of California. All rights reserved.

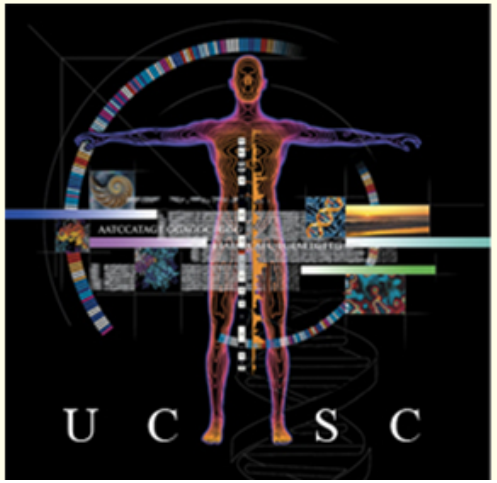
clade	genome	assembly	position or search term	gene	
Mammal	Human	Feb. 2009 (GRCh37/hg19)	chr7:55086725-55275031	EGFR	submit

Click here to reset the browser user interface settings to their defaults.

track search | add custom tracks | track hubs | configure tracks and display | clear position

About the Human Feb. 2009 (GRCh37/hg19) assembly (sequences)

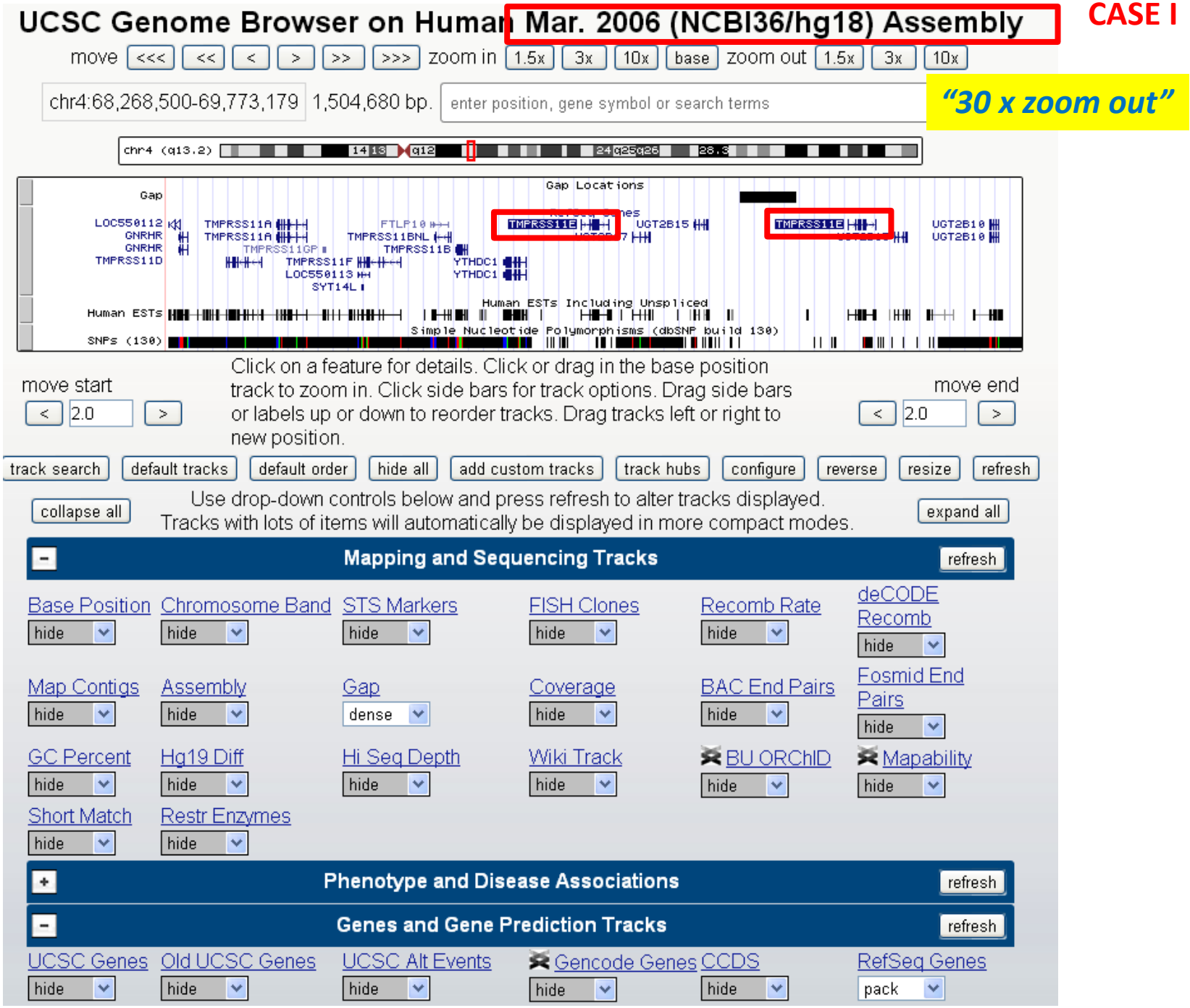
The February 2009 human reference sequence (GRCh37) was produced by the Genome Reference Consortium. For more information about this assembly see GRCh37 in the NCBI Assembly database.



Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the User's Guide for more information.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
chrUn_gl000212	Displays all of the unplaced contig gl000212
20p13	Displays region for band p13 on chr 20

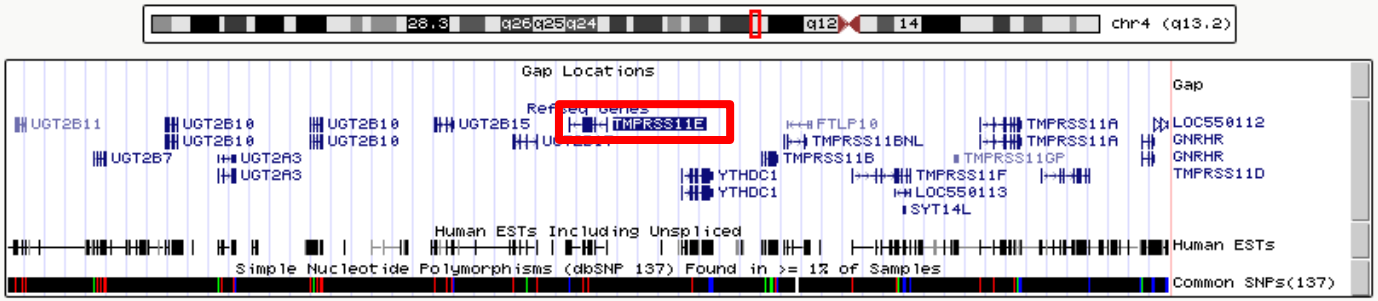

UCSC Genome Browser on Human Mar. 2006 (NCBI36/hg18) Assembly
CASE I
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
chr4:68,268,500-69,773,179 1,504,680 bp.
enter position, gene symbol or search terms
"30 x zoom out"
chr4 (q13.2)
Gap
Gap Locations
LOC550112
GNRHR
GNRHR
TMPRSS11D
TMPRSS11A
TMPRSS11A
TMPRSS11GP
TMPRSS11F
LOC550113
SYT14L
FTLP10
TMPRSS11BNL
TMPRSS11B
YTHDC1
YTHDC1
TMPRSS11E
UGT2B15
TMPRSS11E
UGT2B10
UGT2B10
Human ESTs
Human ESTs Including Unspliced
Simple Nucleotide Polymorphisms (dbSNP build 130)
SNPs (130)
move start
2.0
move end
2.0
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
track search
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh
collapse all
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
expand all
Mapping and Sequencing Tracks
refresh
Base Position hide
Chromosome Band hide
STS Markers hide
FISH Clones hide
Recomb Rate hide
deCODE Recomb hide
Map Contigs hide
Assembly hide
Gap dense
Coverage hide
BAC End Pairs hide
Fosmid End Pairs hide
GC Percent hide
Hg19 Diff hide
Hi Seq Depth hide
Wiki Track hide
BU ORChID hide
Mapability hide
Short Match hide
Restr Enzymes hide
Phenotype and Disease Associations
refresh
Genes and Gene Prediction Tracks
refresh
UCSC Genes hide
Old UCSC Genes hide
UCSC Alt Events hide
Gencode Genes hide
CCDS hide
RefSeq Genes pack

CASE I

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr4:68,585,905-70,090,584 1,504,680 bp. enter position, gene symbol or search terms go



move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

move end < 2.0 >

track search | default tracks | default order | hide all | add custom tracks | track hubs | configure | reverse | resize | refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

− Mapping and Sequencing Tracks refresh

Base Position hide	Chromosome Band hide	STS Markers hide	FISH Clones hide	Recomb Rate hide	deCODE Recomb hide
ENCODE Pilot hide	Map Contigs hide	Assembly hide	GRC Map Contigs hide	Gap dense	BAC End Pairs hide
Fosmid End Pairs hide	GC Percent hide	GRC Patch Release hide	Hg18 Diff hide	GRC Incident hide	Hi Seq Depth hide
Wiki Track hide	BU ORChID hide	Mapability hide	Short Match hide	Restr Enzymes hide	

+ Phenotype and Disease Associations refresh

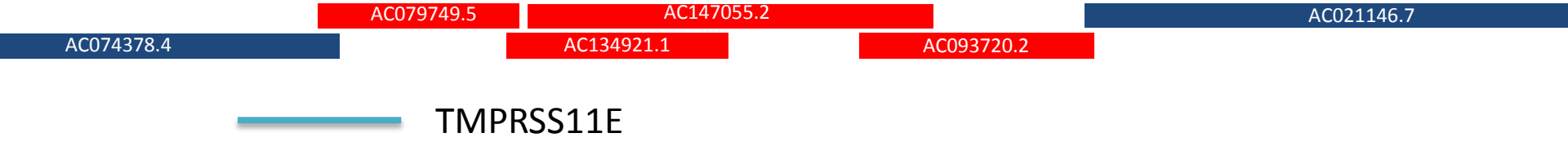
− Genes and Gene Prediction Tracks refresh

UCSC Genes hide	GENCODE... hide	Old UCSC Genes hide	UCSC Alt Events hide	CCDS hide	RefSeq Genes pack

CASE I

NCBI36 NC_000004.10 (chr4) Tiling Path

GRCh37 NC_000004.11 (chr4) Tiling Path



GRCh37: NT_167250.1 (UGT2B17 alternate locus)

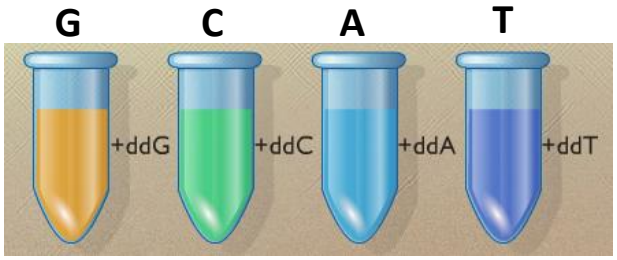
Xue Y et al, 2008

Sequencing

- Basic molecular biology
- Sanger method
- HT sequencing
- Human genome sequencing strategies
  - Directed (Top-down) vs. Shotgun (Bottom-up)
- Next-Generation (Massive Parallel) Sequencing
  - Pyrosequencing (Roche/454)
  - Sequencing by synthesis (Illumina)
  - Sequencing by ligation (SOLiD, AB)

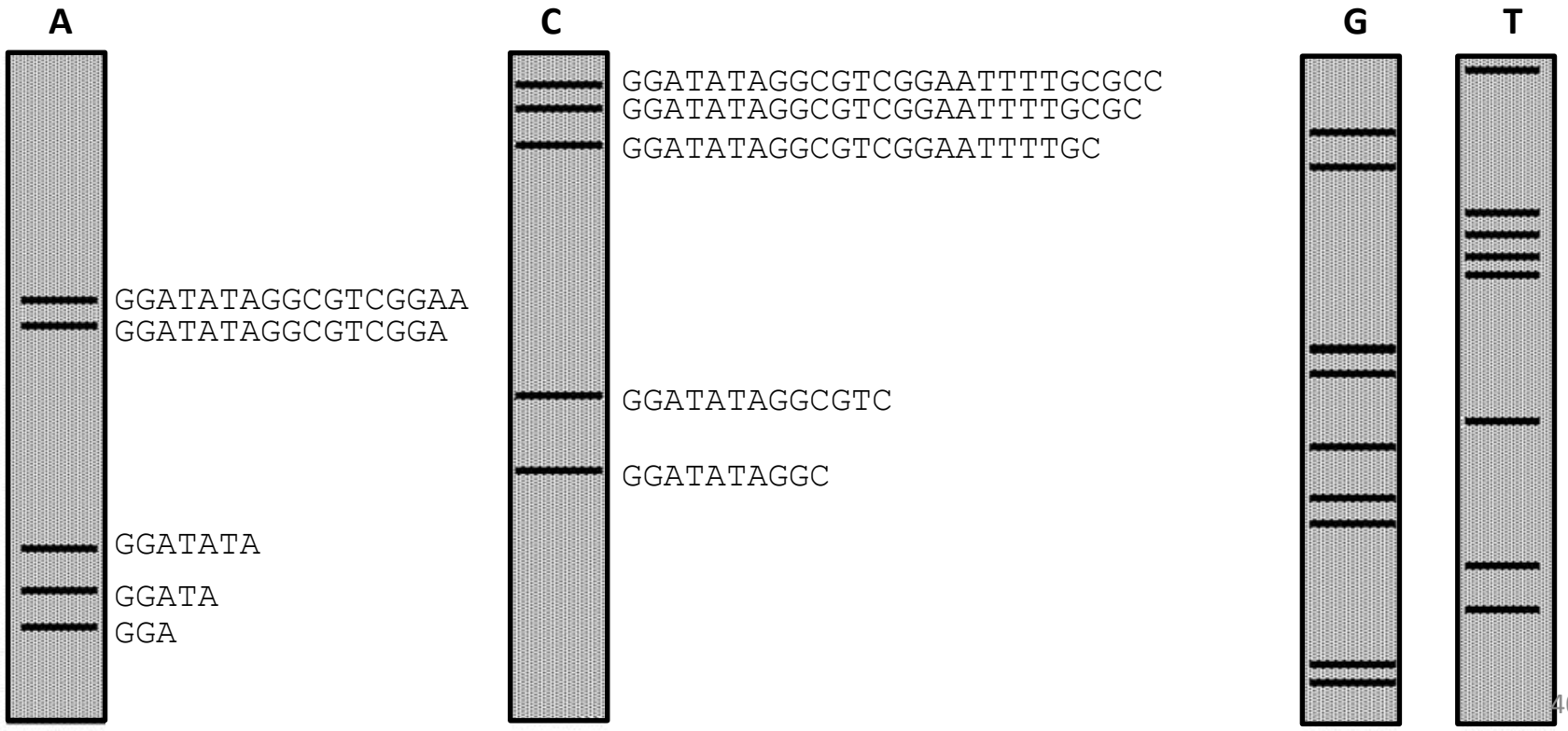
Generation I: Manual Sequencing

Sanger Method: Chain Termination Method



A: all four dNTP's, **ddATP** and DNA polymerase
C : all four dNTP's, **ddCTP** and DNA polymerase
G: all four dNTP's, **ddGTP** and DNA polymerase
T : all four dNTP's, **ddTTP** and DNA polymerase

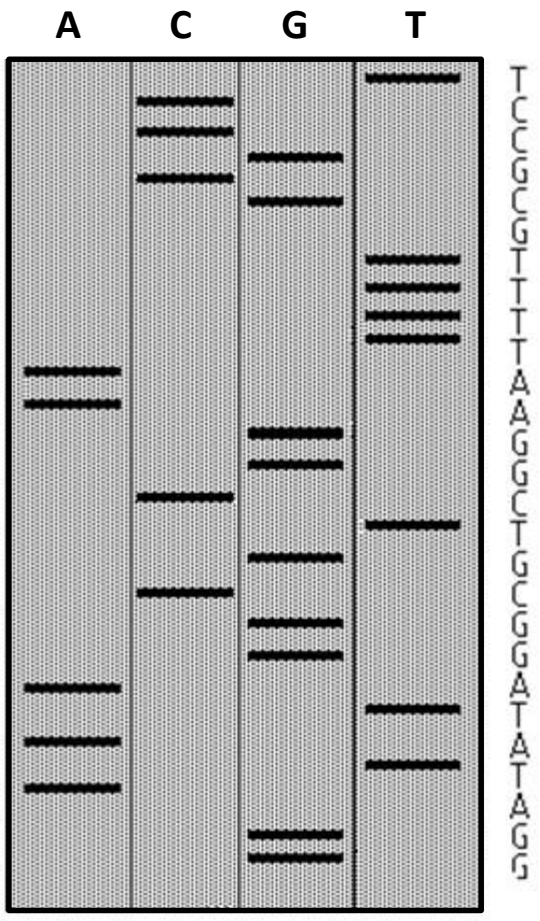
GGATATAGGCGTCGGAATTTTGCGCCT

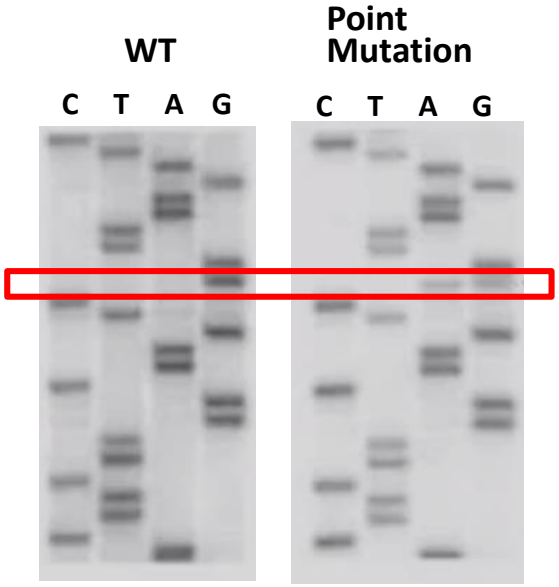


Generation I: Manual Sequencing

Sanger Method: Chain Termination Method

Composition



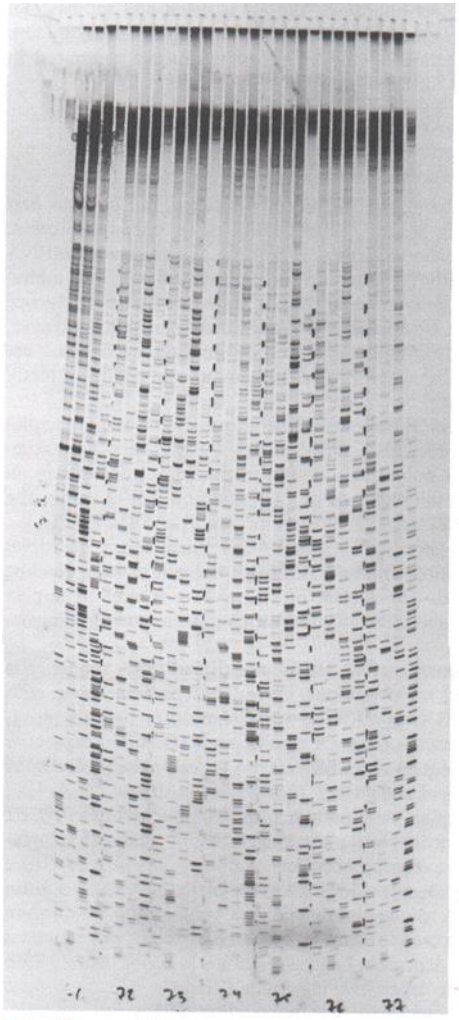


Radioisope labeld ddNTP

Loading each ddNTP reaction in a different lane

Generation I: Manual Sequencing

Sanger Method: Chain Termination Method



Generation II: Automated Sequencing

http://www.genomenewsnetwork.org/resources/timeline/

Genetics and Genomics Timeline

ShareThis

Printer Friendly

1986

Leroy Hood (1938-) develops the automated sequencer

The techniques for sequencing DNA, developed independently by Walter Gilbert and Frederick Sanger in the late 1970s, represented dramatic advances for research in molecular biology. But they were labor intensive and costly. Automation was always a goal, and just about the time that scientists began to confront the prospect of sequencing the human genome, it became a reality. In 1986, Applied Biosystems Incorporated (API) announced the first automatic sequencer, invented by Leroy Hood.

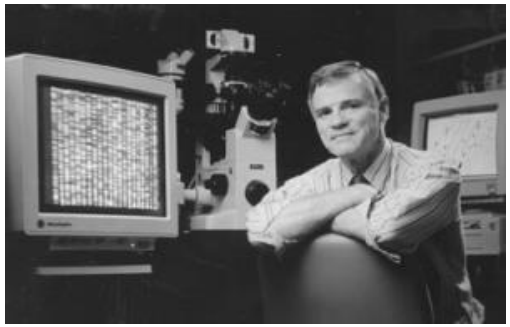
A biologist at the California Institute of Technology and a founder of API, Hood improved the existing Sanger' method of enzymatic sequencing, which was becoming the laboratory standard. In this method, DNA to be sequenced is cut apart, and a single strand serves as a template for the synthesis of complementary strands. The nucleotides used to build these strands are randomly mixed with a radioactively labeled and modified nucleotide that terminates the synthesis. Fragments of all different lengths result. The resulting array, sent through a separation gel, reveals the order of the bases. Transferred to film, an "autoradiograph" provides a readable sequence from raw data. This data could be transferred to a computer by a human reader.

In automating the process, Hood modified both the chemistry and the data-gathering processes. In the sequencing reaction itself, he sought to replace the use of radioactive labels, which were unstable, posed a health hazard, and required separate gels for each of the four DNA bases.

• In place of radioisotopes, Hood developed chemistry that used fluorescent dyes of different colors—one for each of the four DNA bases. This system of "color-coding" eliminated the need to run several reactions in overlapping gels.

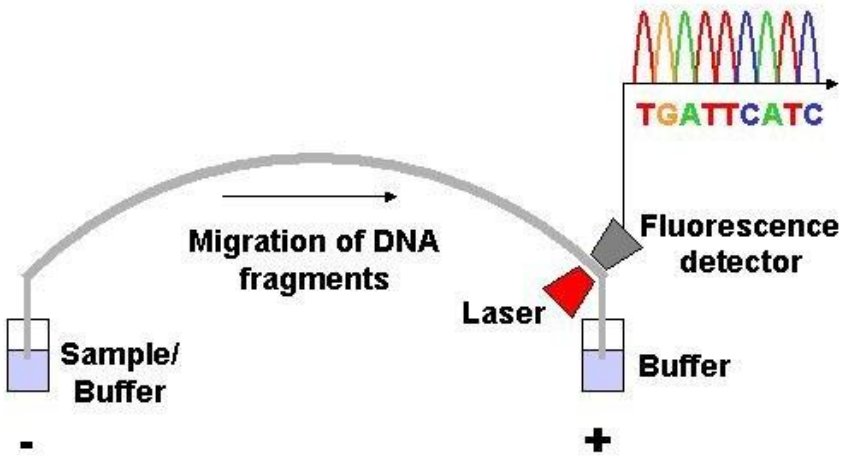
The fluorescent labels were also aspects of the larger system that revolutionized the end stage of the process—the way in which sequence data was gathered. Hood integrated laser and computer technology, eliminating the tedious process of information-gathering by hand.

• As the fragments of DNA percolated through the gel, a laser beam stimulated the fluorescent labels, causing them to glow. The light they emitted was picked up by a lens and photomultiplier, and transmitted as digital information directly into a computer.

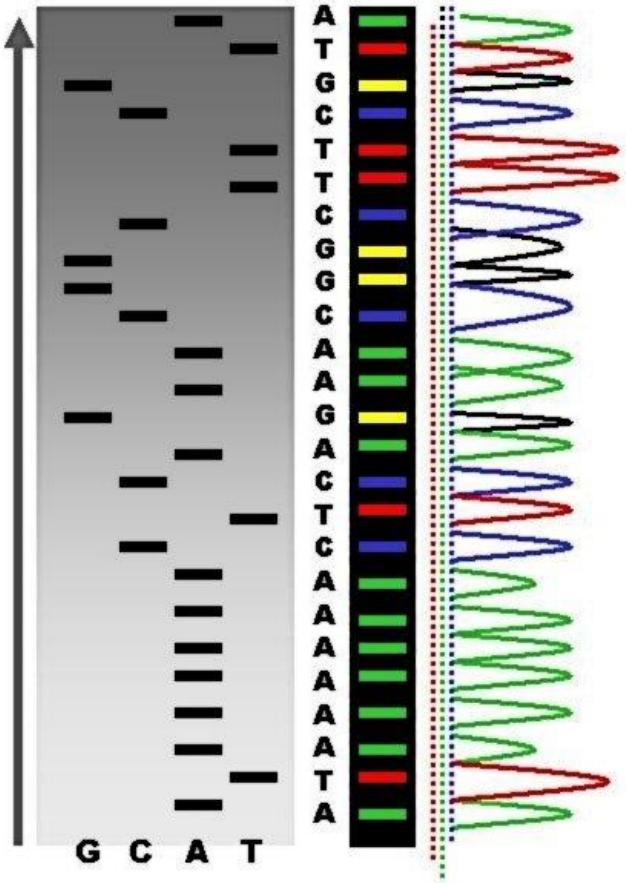


- **Fluorescent dye labeled ddNTP**
- **HT Data handling by integrated laser/computer technology**

Generation II: Automated Sequencing



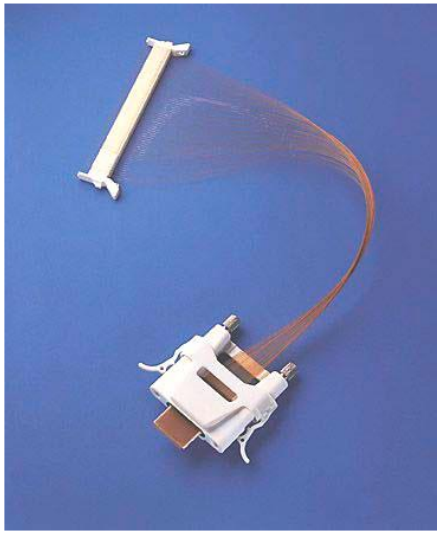
Automated sequencing machines use four colors, so they can read all four bases at once.

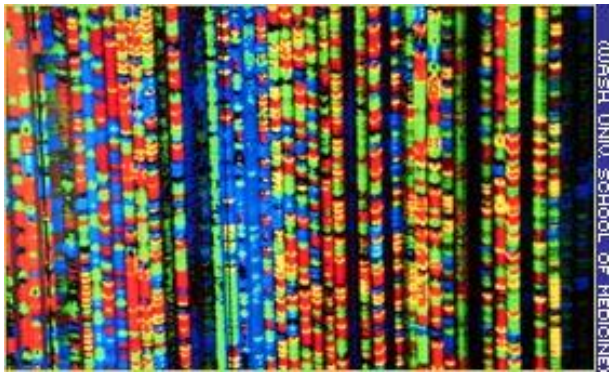


Sequence ladder by radioactive sequencing compared to fluorescent peaks

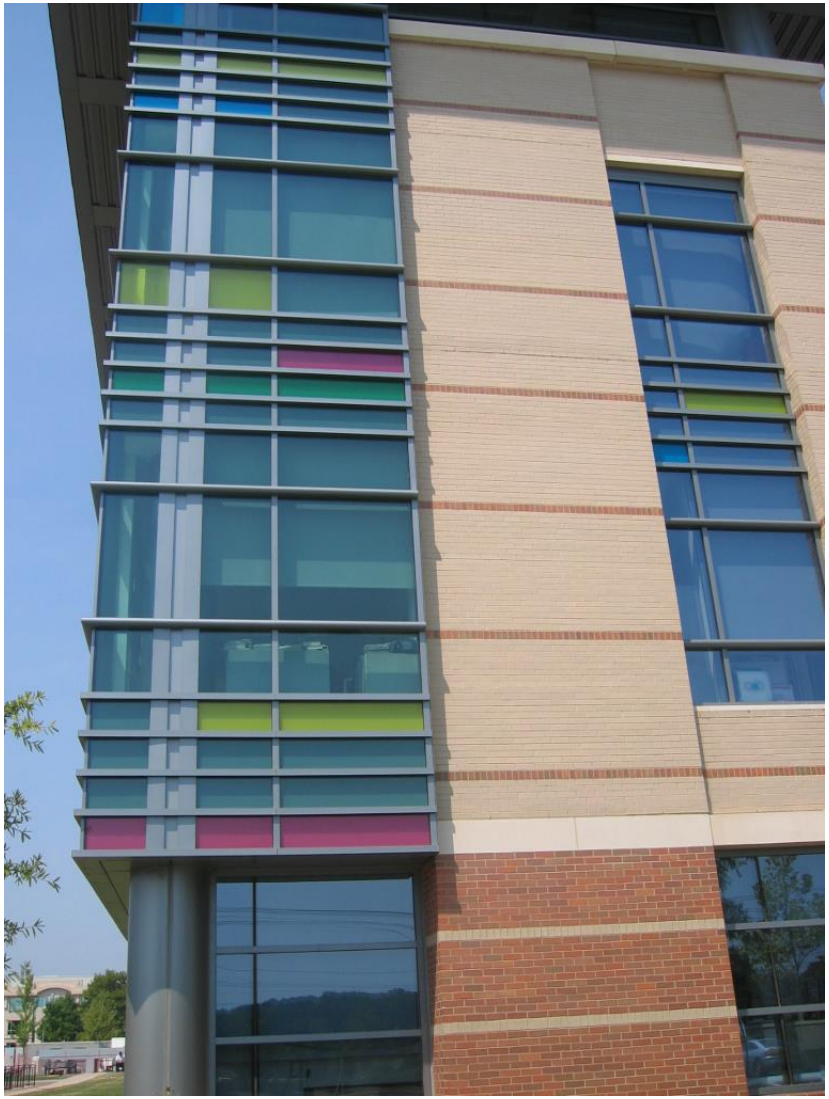
Generation II: Automated Sequencing

ABI 3730xl





Generation II: Automated Sequencing

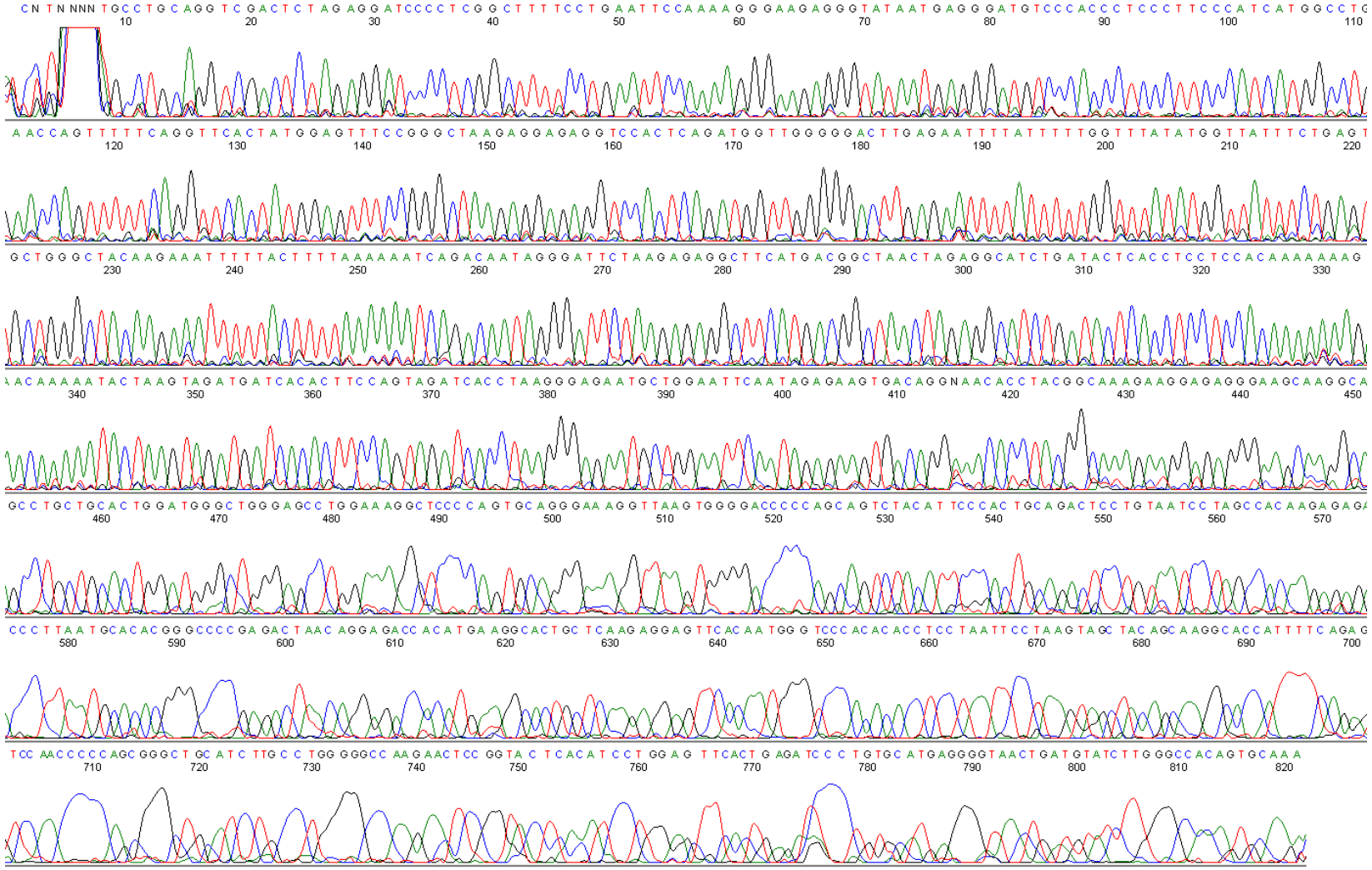


AGCT

JC Venter Institute, Rockville MD

Generation II: Automated Sequencing

Trace File



Exercise

- http://www.clcbio.com/index.php?id=92
- Download free trial
- Transfer sequencing data

CASE II. The provided two data sets contain CE based sequencing information from ABI platform. The first set contains different fragments of DNAs from one patient, the second set contains PCR products of the same region of specific gene from multiple patients. Import the data set to your computer, and analyze the data accordingly (1) to identify the region of sequences in reference genome, and (2) to find patients who have variants and the meaning of the variants.

Exercise

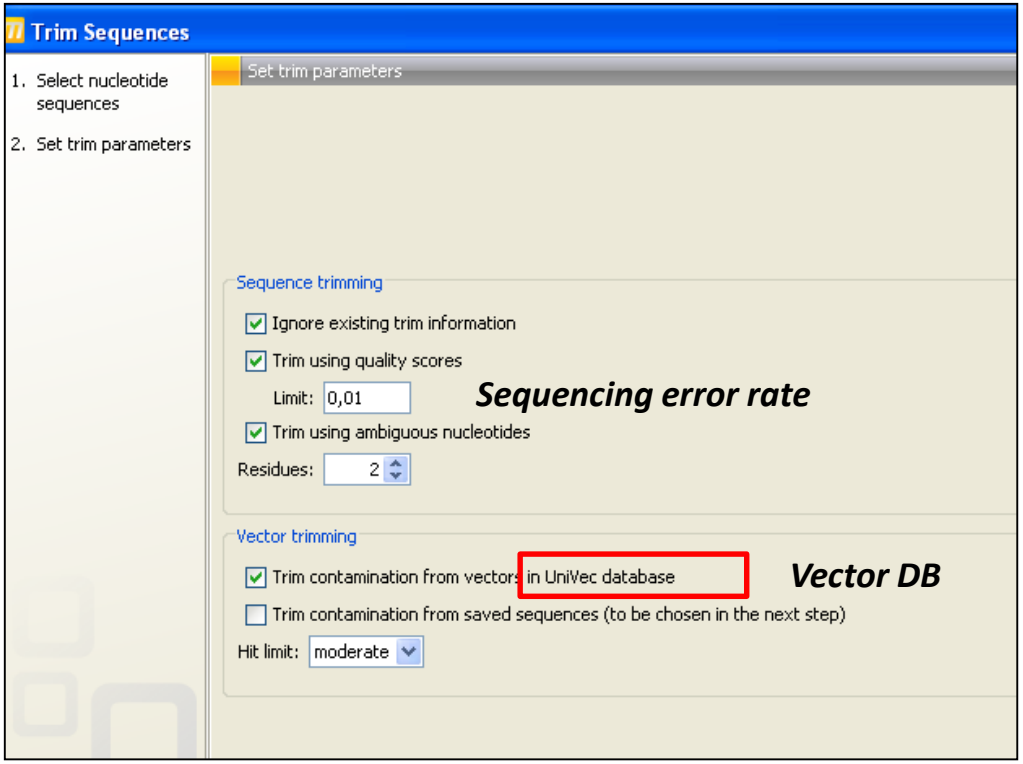
<Part1>

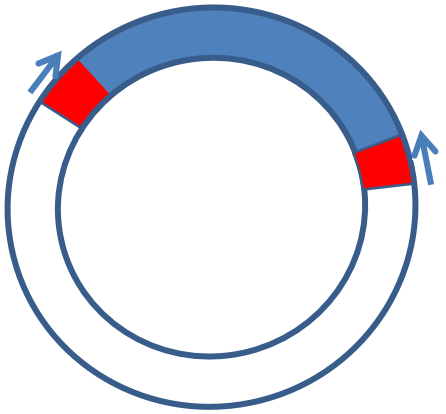
1. Assemble the gene fragments derived from a patient.
2. Identify the gene using BLASTn (RefSeq mRNA).
3. Download the sequence file of identified gene from NCBI.
4. Map the contigs to the reference sequence.

<Part2>

5. Align the PCR products from multiple patients.
6. Find the conflict sequences.
7. BLAT at the genome browser to map the variant in the reference and look for annotations
8. Find an associated SNP track
9. SNP DB-> rs number
10. NCBI -> Gene information
11. UniProt/SwissProt -> Protein information, look for variant annotation
12. Disease association

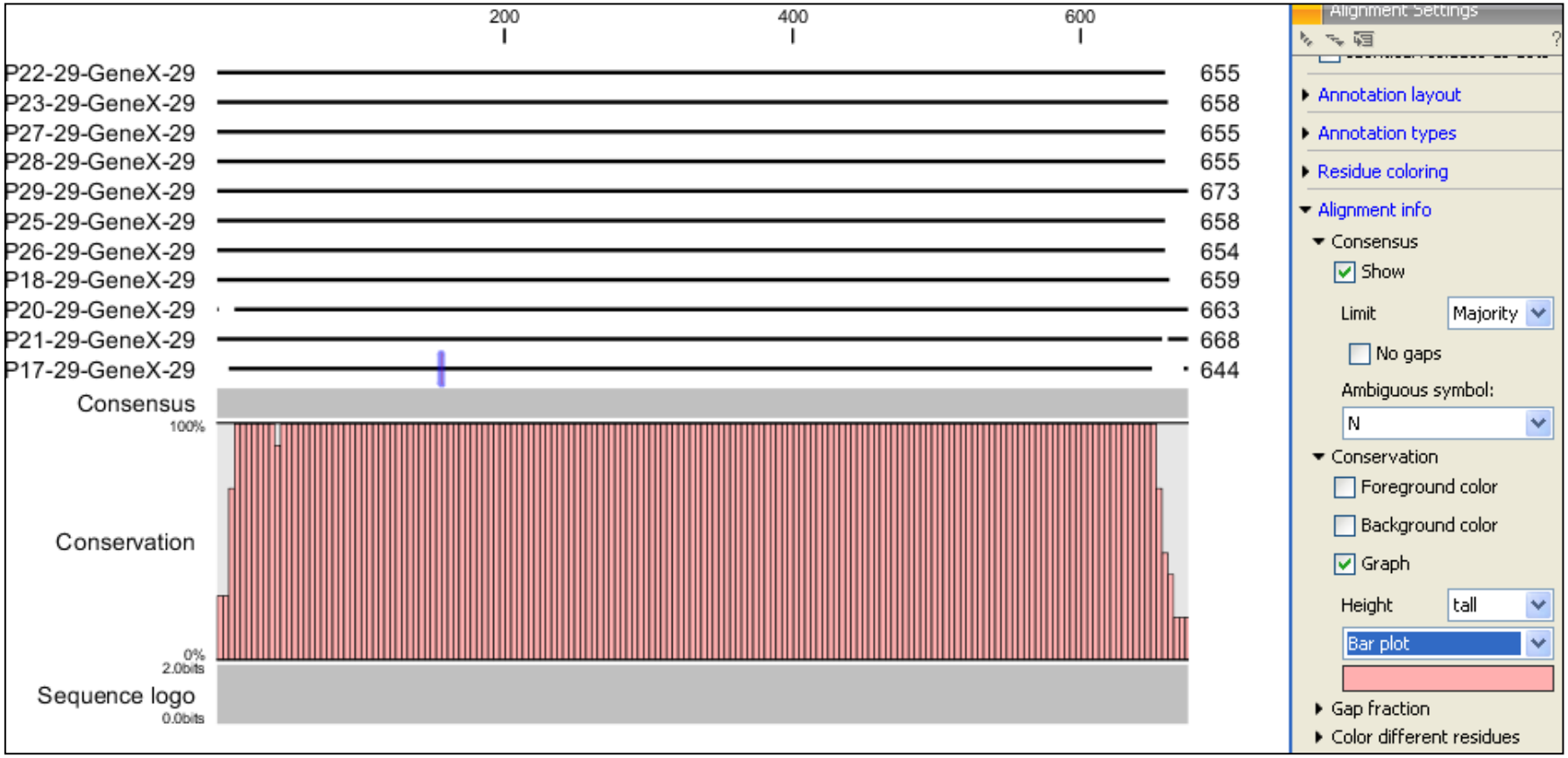
Why trimming?





CASE II

Find SNP



BLAT search and linkage of DBs

BLAT in Genome Browser using GRCh37

Gene Name:

SNP ID:

Synonymous or Non-synonymous:

Disease Association:

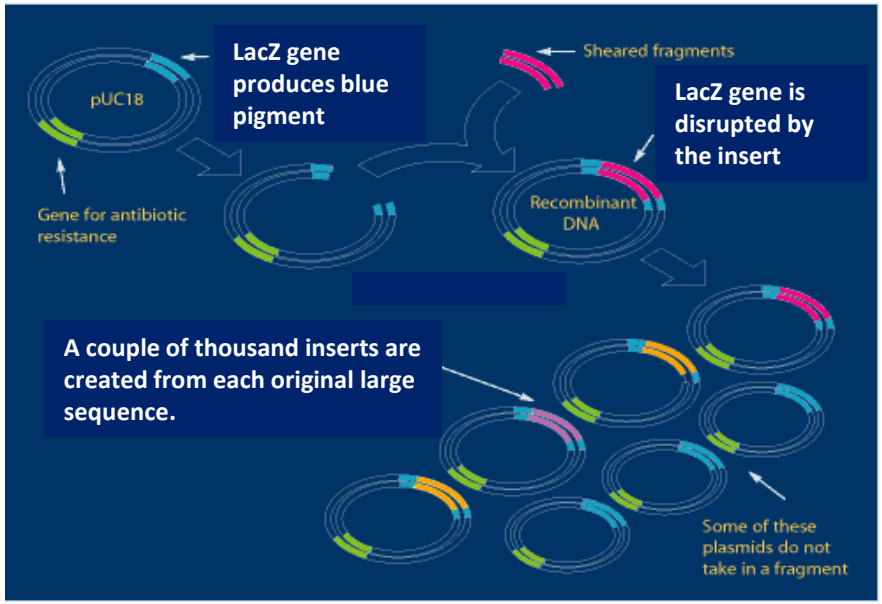
Find RefSeq Protein accession number:

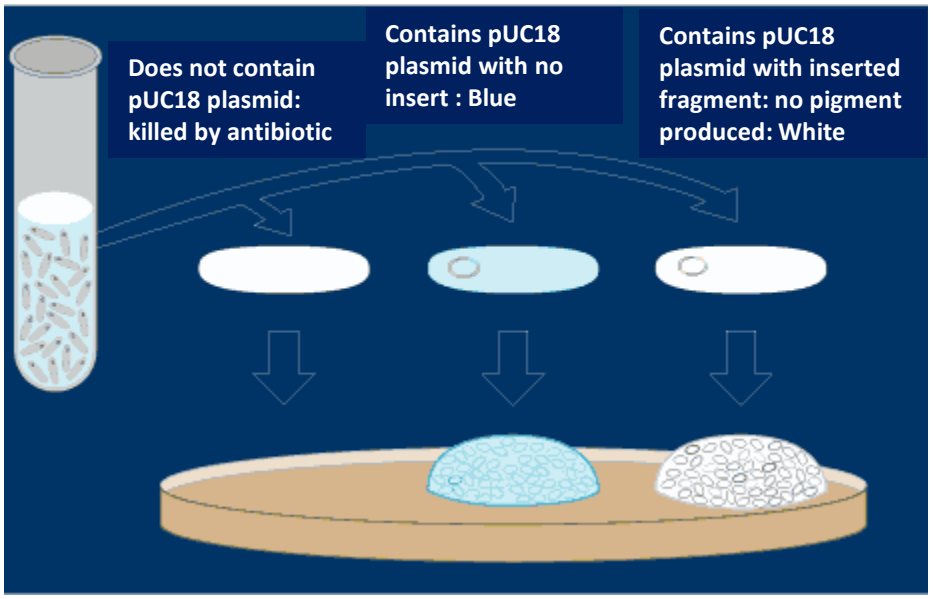
Search the AN in UniProt:

Review and find the SNP ID:

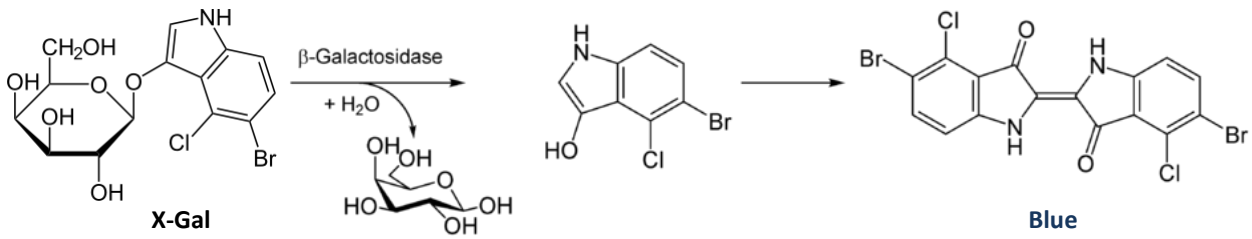
Check reference

Cloning was necessary for prior to G-III

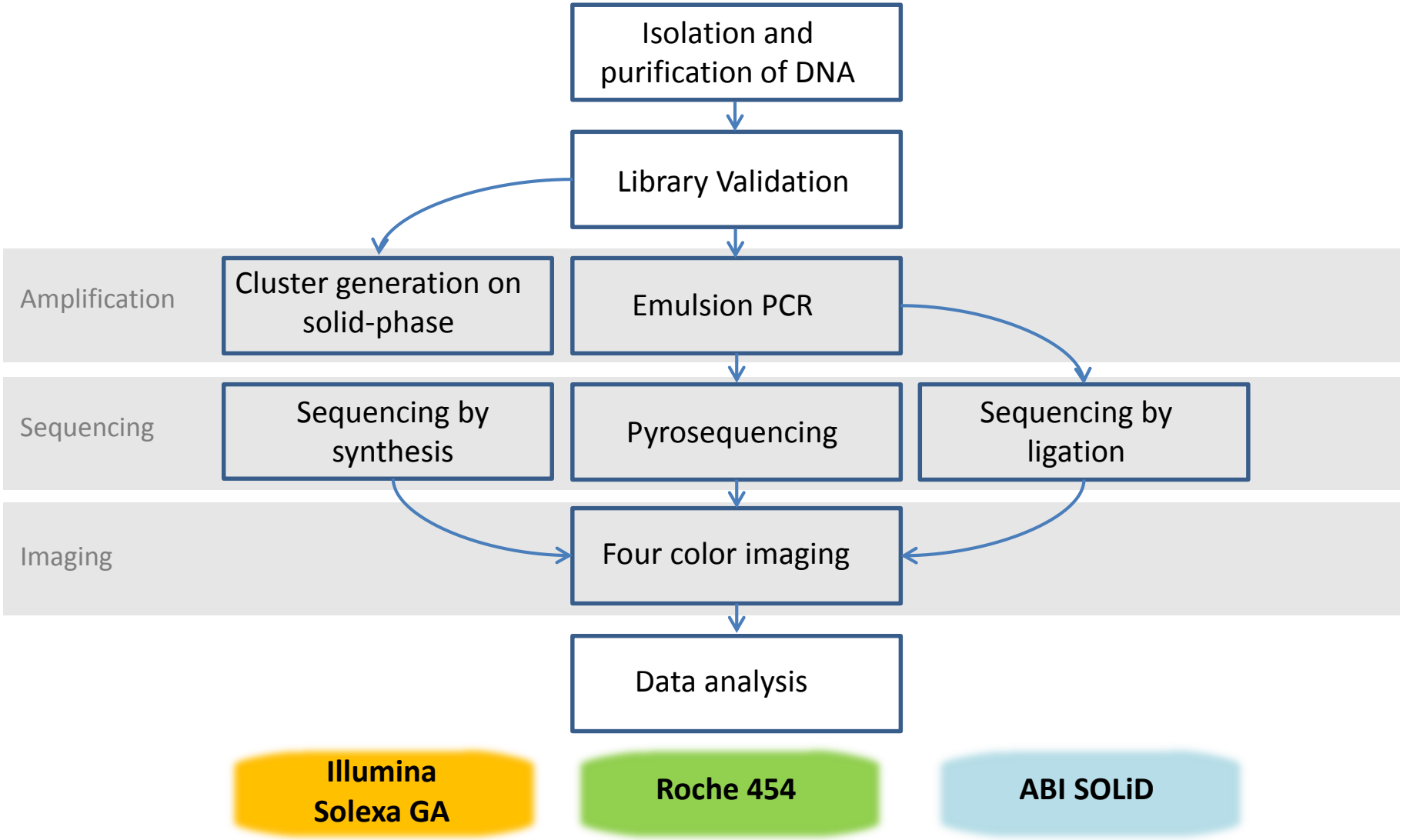




β-Galactosidase Exist **β-Galactosidase Absent**



Generation III: Solid Phase Clusters
(Massively Parallel Sequencing)



Slide adapted from http://www.slideshare.net/cosentia/high-throughput-equencing

Generation III: Solid Phase Clusters
(Massively Parallel Sequencing)

Illumina Solexa GA	Roche 454	ABI SOLiD
Pyrosequencing	Seq by Synthesis	Seq by Ligation

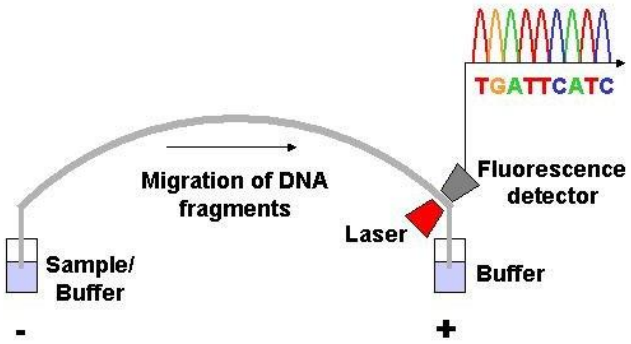
Questions:

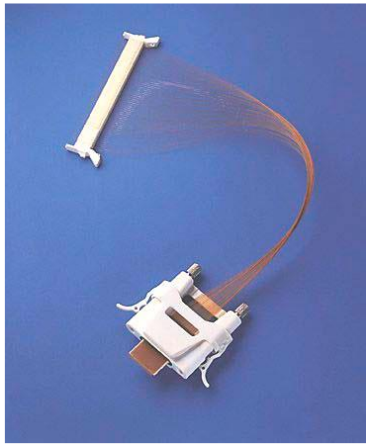
1. Major differences of the 2nd Generation Sequencing (eg ABI3730) from classic Sanger Methods.

2. Major differences of NGS from the 2nd GS.

Throughput

Speed is matter





If 500b read per capillary
500 x 100 b =50,000b per 3 hr
50,000b x 8 = 400,000b per 1 day

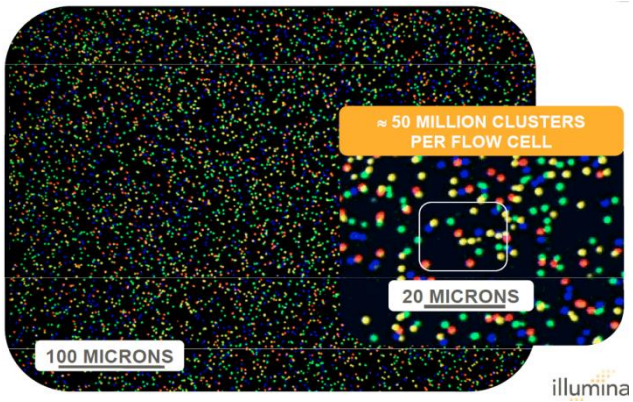
If 1000b read per capillary
800,000b per 1day

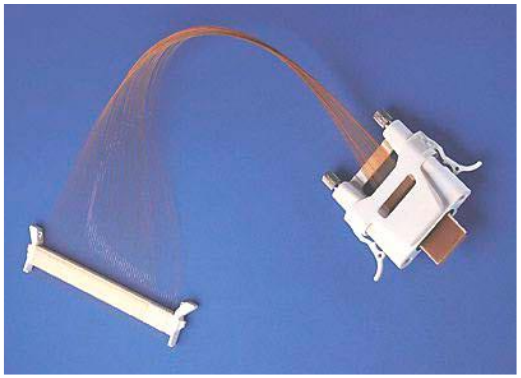
For whole genome
$3x10^9$ b -> 3,750 d -> 10.3 y

Comparison of sequencers

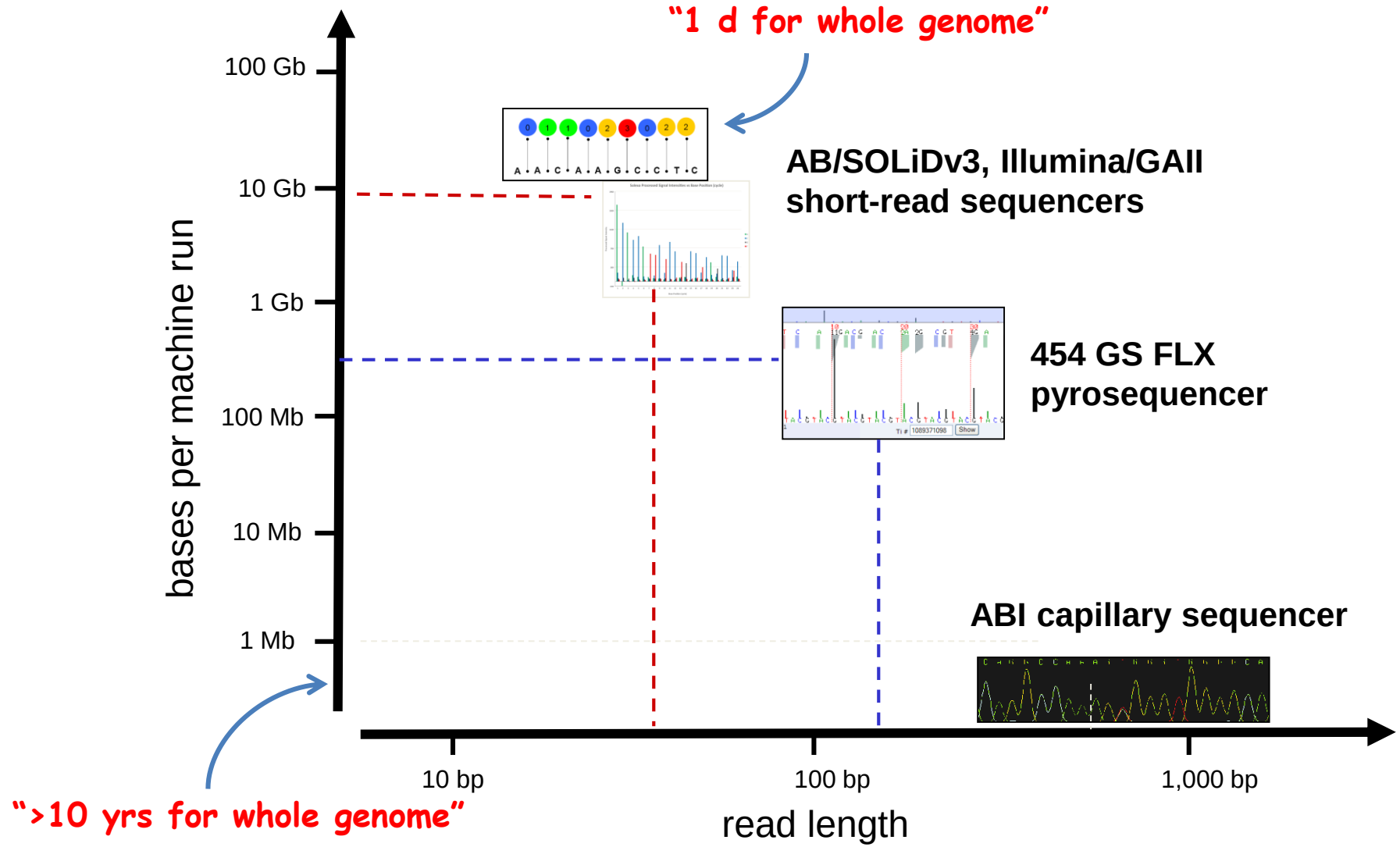
Vendor	Roche			Illumina			ABI			ABI
Technology	454			Solexa GA			SOLiD			3730
Platform	GS20	FLX	Ti	I	II	IIx	1	2	3	
Reads (M)	0.5	0.5	1.25	28	100	250	40	115	320	9.60E-05
Fragment										
Read length	100	200	400	35	50	100	25	35	50	800
Run time (d)	0.25	0.3	0.4	3.0	3.0	5.0	6.0	5.0	8.0	0.125
Yield (Gb)	0.05	0.1	0.5	1.00	5.00	25.00	1.00	4.00	16.00	7.68E-05
Rate (Gb/d)	0.2	0.33	1.25	0.33	1.67	5.00	1.67	0.8	2.00	6.14E-04
For 3Gb	**15.0**	**9.1**	**2.4**	**9.09**	**1.8**	**0.6**	**1.8**	**3.8**	**1.5**	**4882.8**

http://www.politigenomics.com/next-generation-sequencing-informatics





Comparison of sequencers



The 1000 Genomes project

Nature method, 2012

The 1000 Genomes Project was launched as one of the largest distributed data collection and analysis projects ever undertaken in biology. In addition to the primary scientific goals of creating both a deep catalog of human genetic variation and extensive methods to accurately discover and characterize variation using new sequencing technologies, the project makes all of its data publicly available. Members of the project data coordination center have developed and deployed several tools to enable widespread data access.

High-throughput sequencing technologies, including those from Illumina, Roche Diagnostics (454) and Life Technologies (SOLiD), enable whole-genome sequencing at an unprecedented scale and at dramatically reduced costs over the gel capillary technology used in the human genome project. These technologies were at the heart of the decision in 2007 to launch the 1000 Genomes Project, an effort to comprehensively characterize human variation in multiple populations. In the pilot phase of the project, the data helped create an extensive population-scale view of human genetic variation[1].

The larger data volumes and shorter read lengths of high-throughput sequencing technologies created substantial new requirements for bioinformatics, analysis and data-distribution methods. The initial plan for the 1000 Genomes Project was to collect 2× whole genome coverage for 1,000 individuals, representing ~6 giga–base pairs of sequence per individual and ~6 tera–base pairs (Tbp) of sequence in total. Increasing ~40× whole-genome sequence for 500 individuals in total (~25-fold increase in sequence generation over original estimates). In fact, the 1000 Genomes Pilot Project collected 5 Tbp of sequence data, resulting in 38,000 files and over 12 terabytes of data being available to the community[1]. In March 2012 the still-growing project resources include more than 260 terabytes of data in more than 250,000 publicly accessible files.

As in previous efforts[2–4], the 1000 Genomes Project members recognized that data coordination would be critical to move forward productively and to ensure the data were available to the community in a reasonable time frame. Therefore, the Data Coordination Center (DCC) was set up jointly between the European Bioinformatics Institute (EBI) and the National Center for Biotechnology (NCBI) to manage project-specific data flow, to ensure archival sequence data deposition and to manage community access through the FTP site and genome browser.

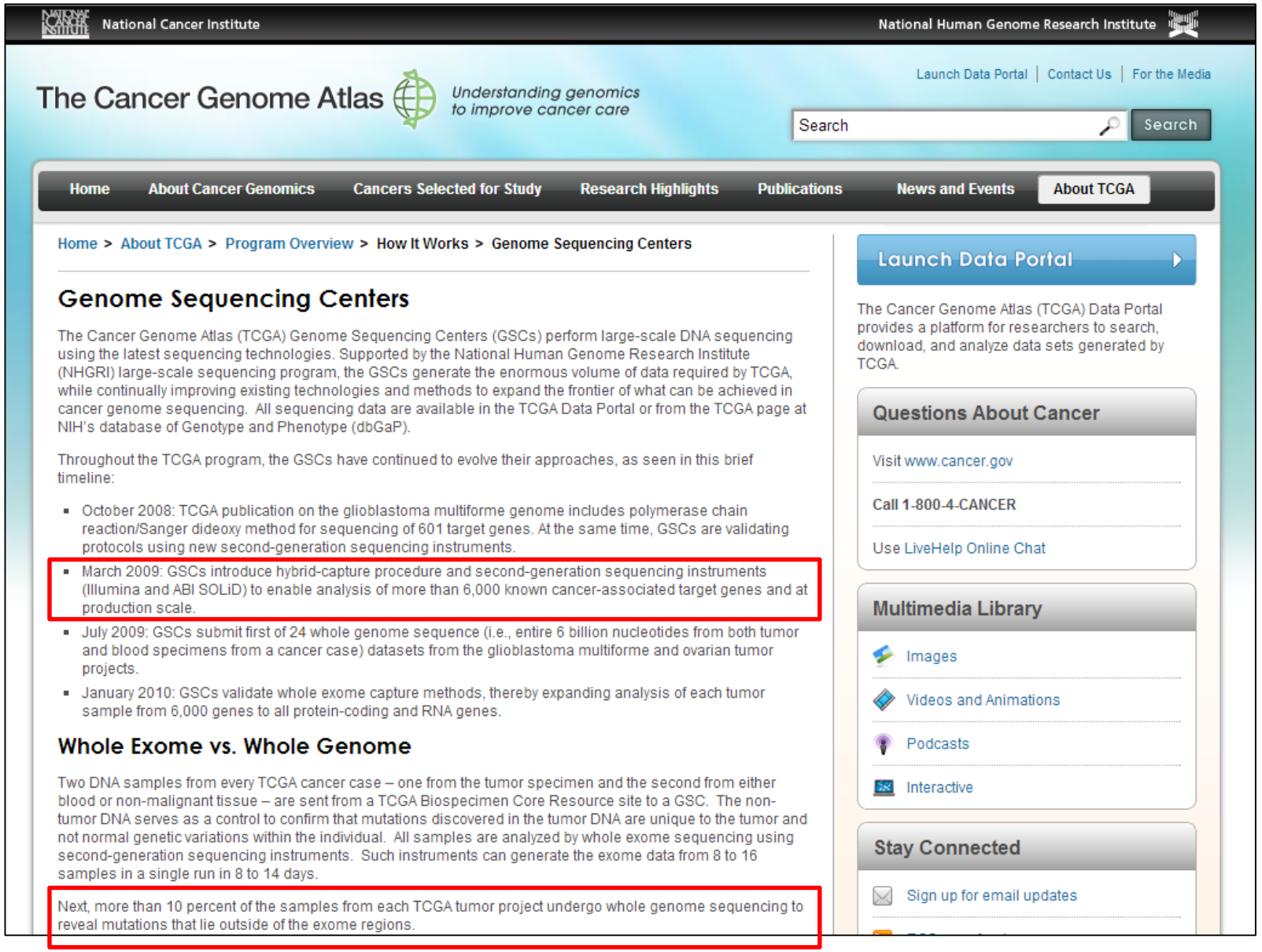
Here we describe the methods used by the 1000 Genomes Project members to provide data resources to the community from raw sequence data to project results that can be browsed. We provide examples drawn from the project's data-processing methods to demonstrate the key components of complex workflows.

Data flow

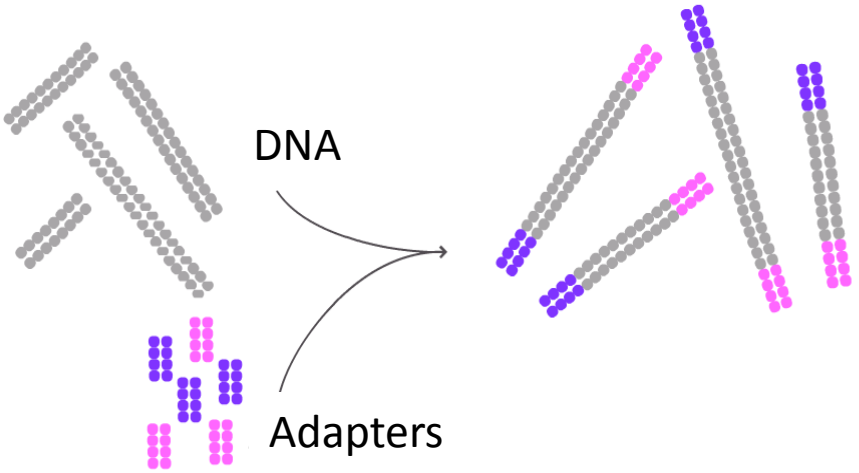
Managing data flow in the 1000 Genomes Project such that the data are available within the project and to the wider community is the fundamental bioinformatics

Check out this! http://genomics.xprize.org/

The Cancer Genome Atlas

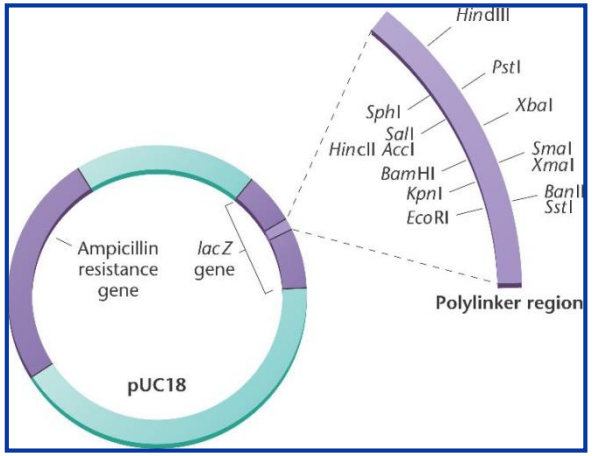


In NGS, adapters for DNA preparation



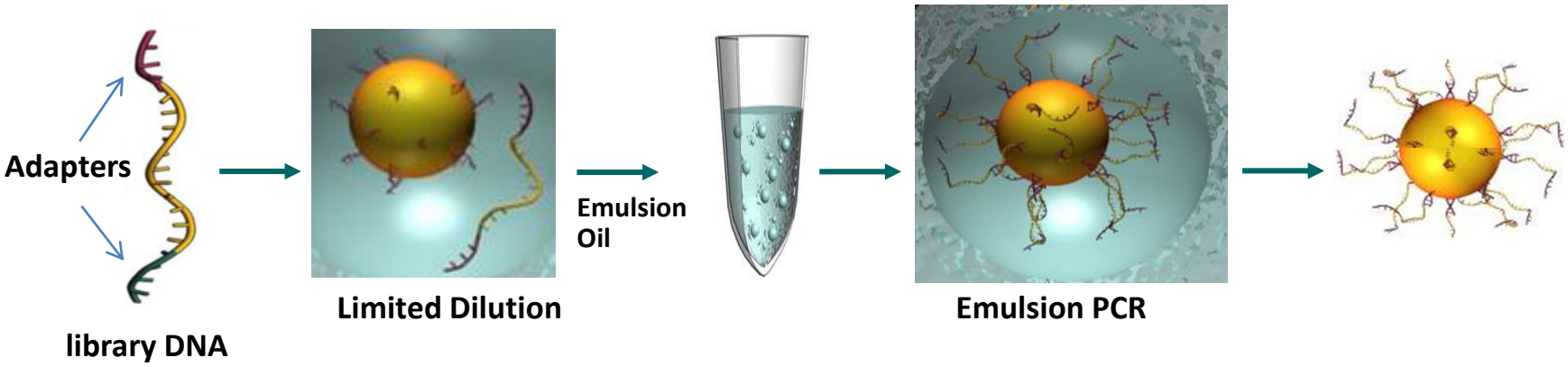
Roles of Adapters : Primer
Immobilizer
Barcode

instead of this

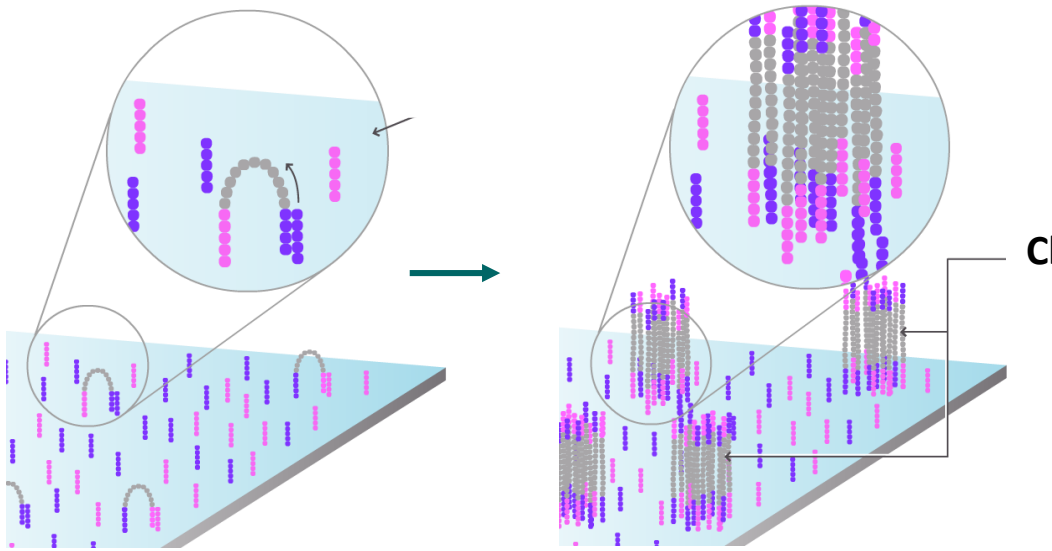


In NGS, instead of CLONING!

1. Emulsion PCR (Roche, AB)



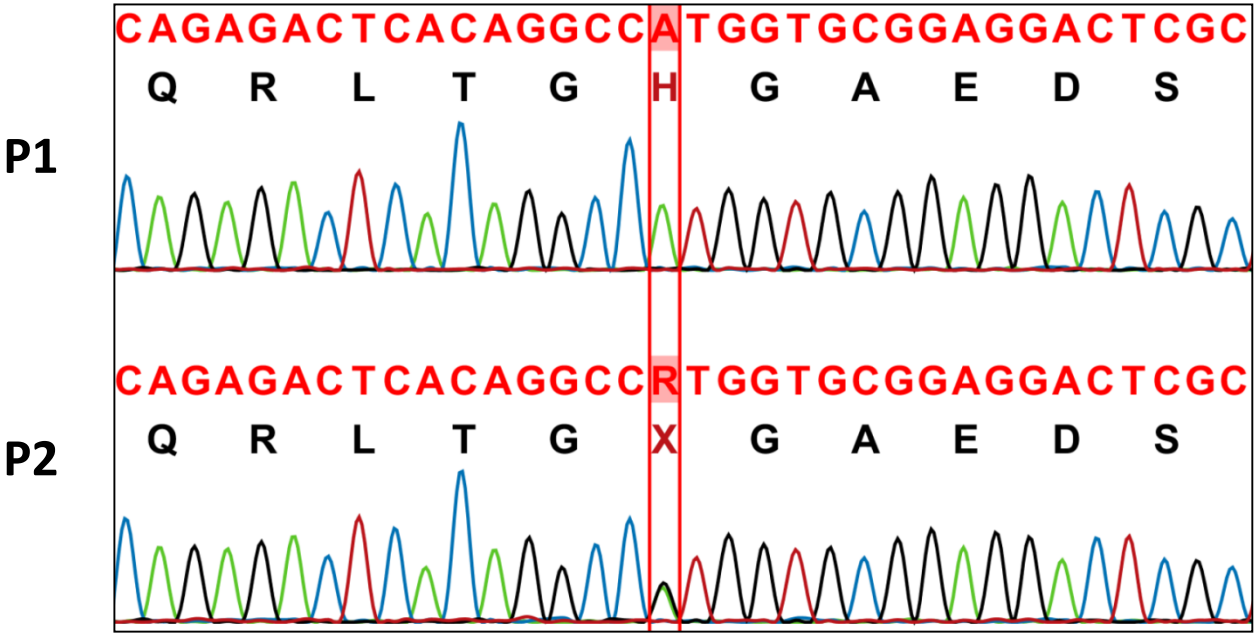
2. Bridge Amplification (Illumina)



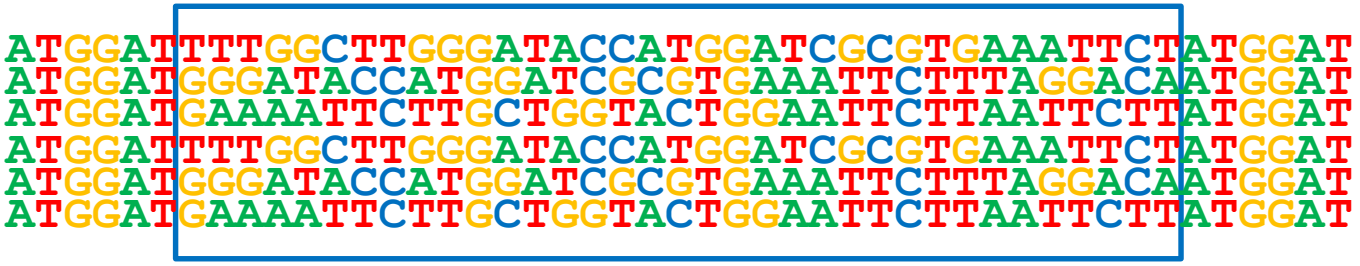
"Insertion 700b"

How to handle these cases in G-II vs G-III?

(1)

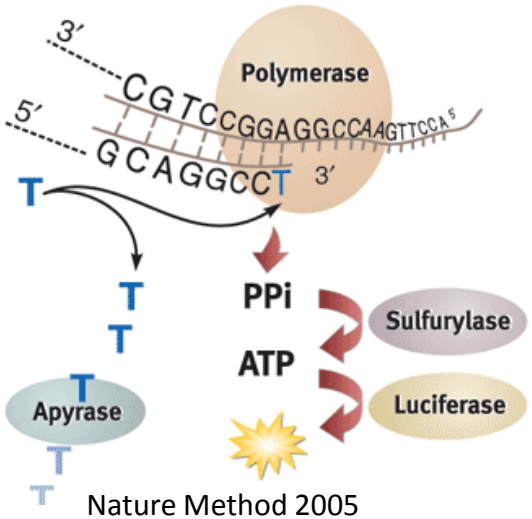


(2)

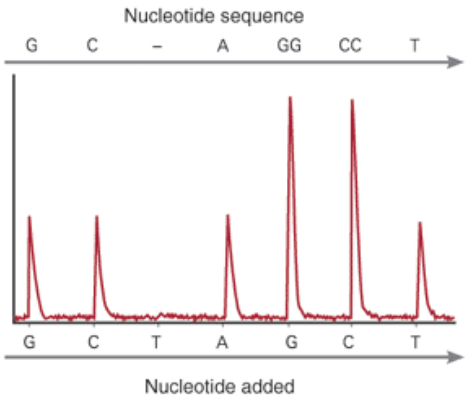


In NGS, for Base Calling

1. Pyrosequencing (Roche)



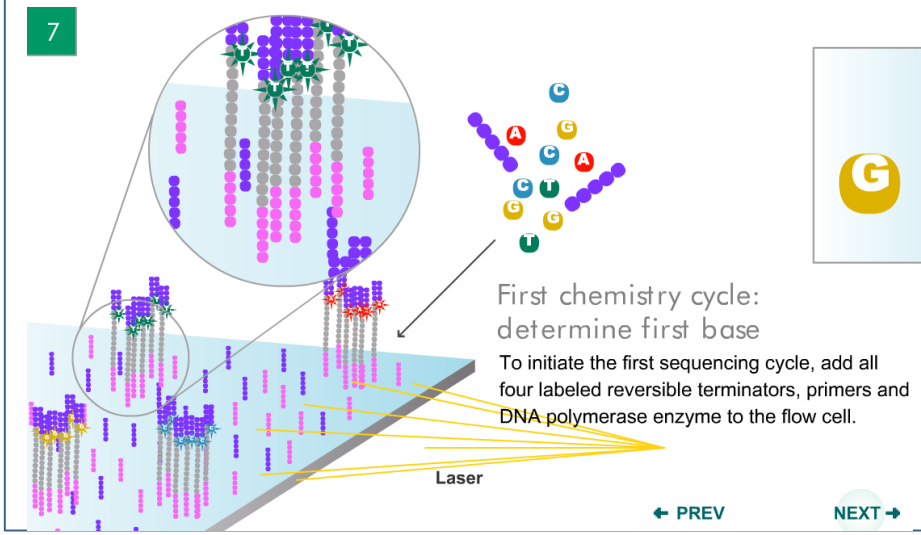
Nature Method 2005

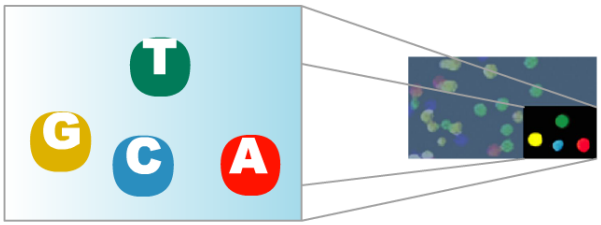


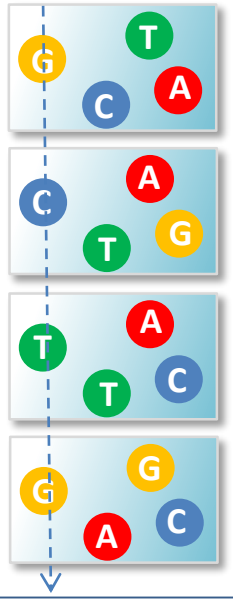
Excessive use of dNTPs, Enzymes (Sulfurylase, Luciferase, Plymerase, Apyrase)

Weak: Homopolymeric repeats

2. Seq by synthesis (Illumina)

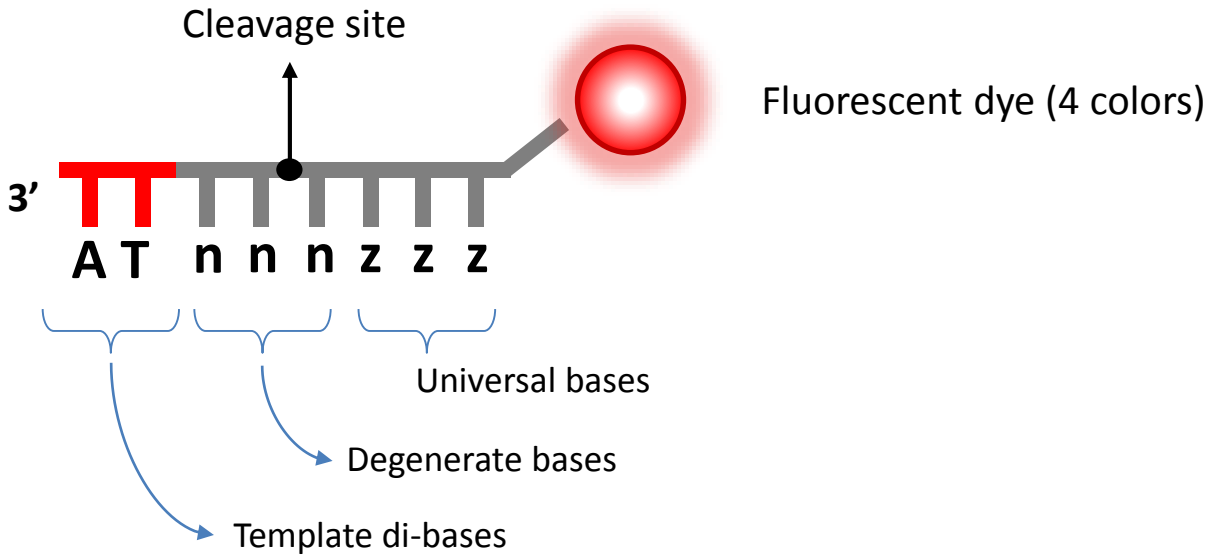






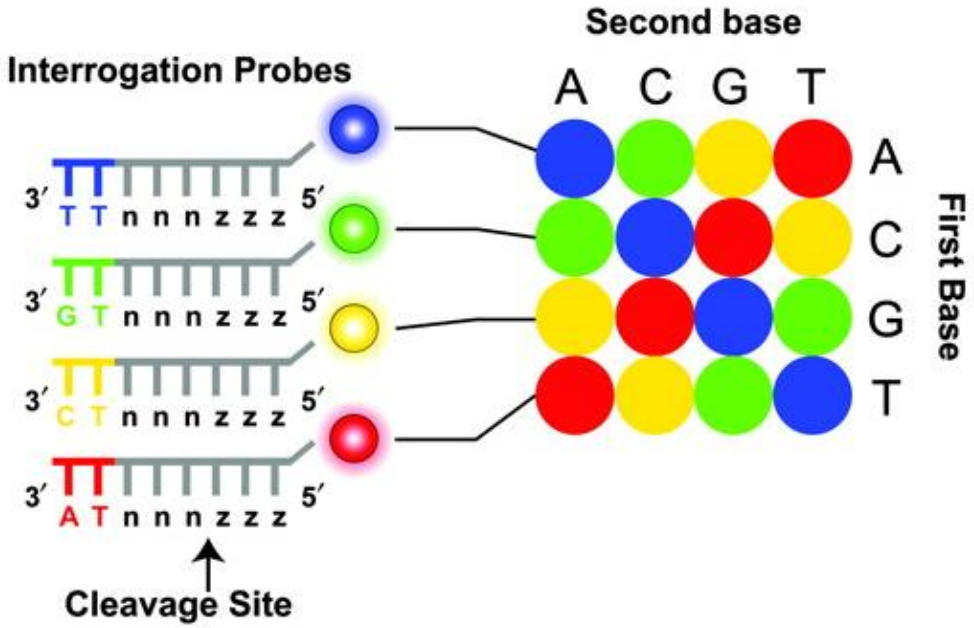
In NGS, for Base Calling

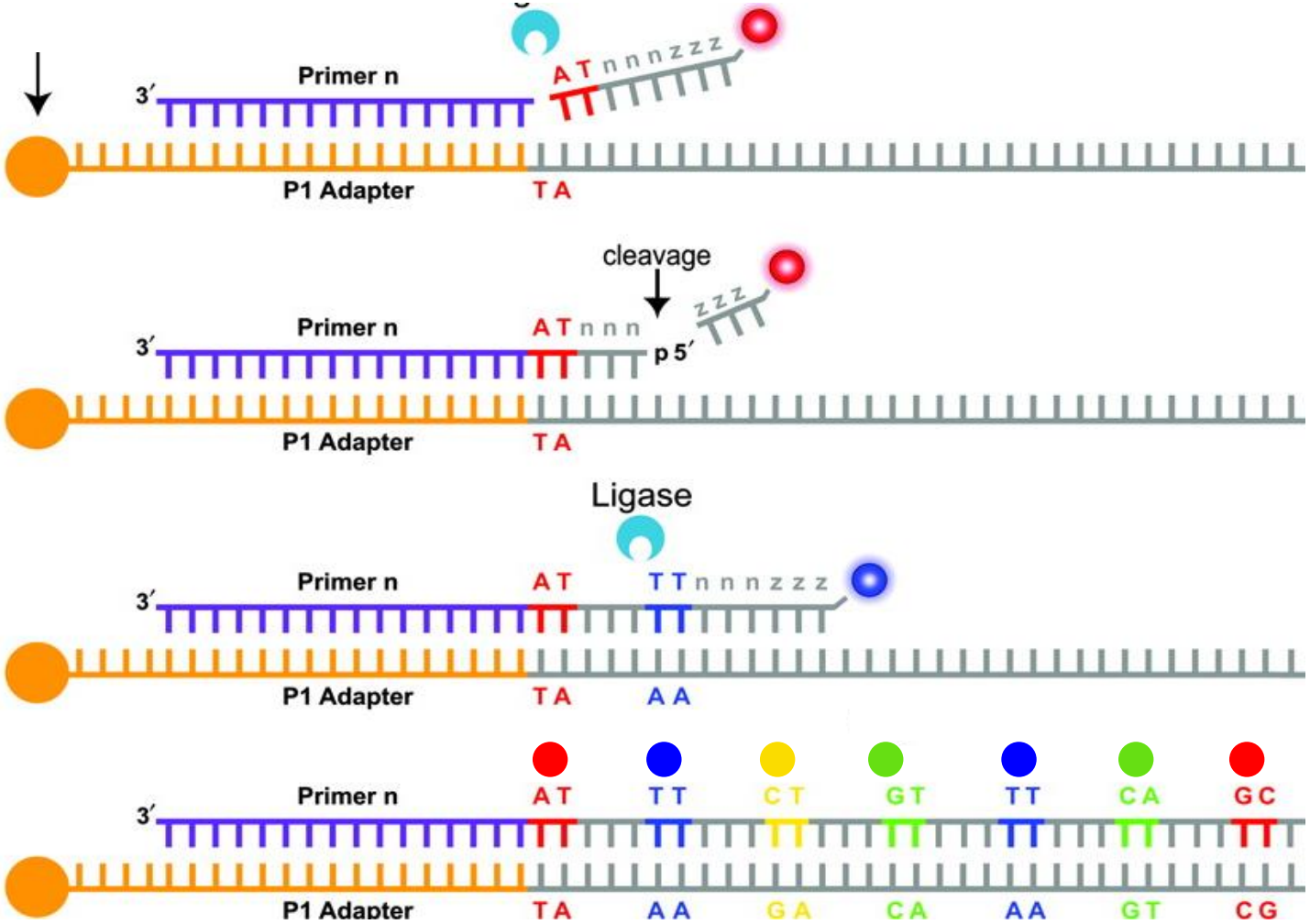
(Special attention to SOLiD Platform)

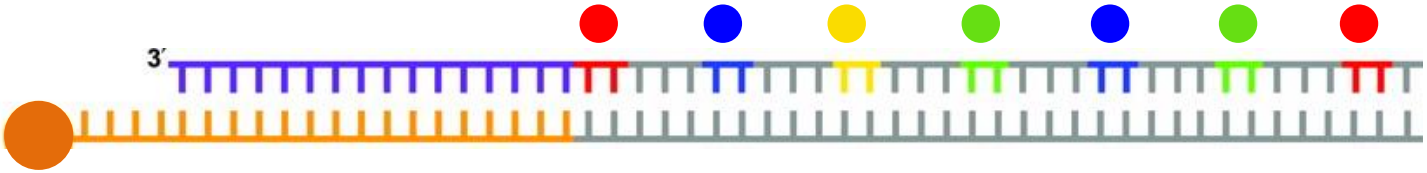


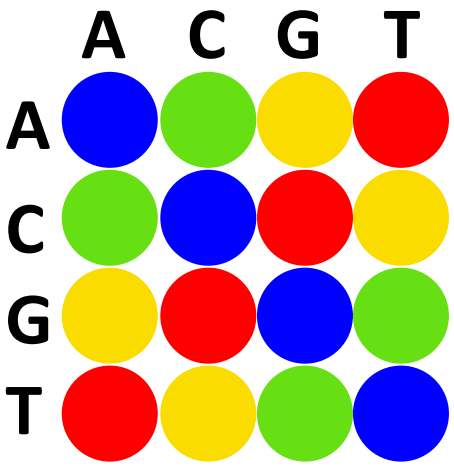
In NGS, for Base Calling

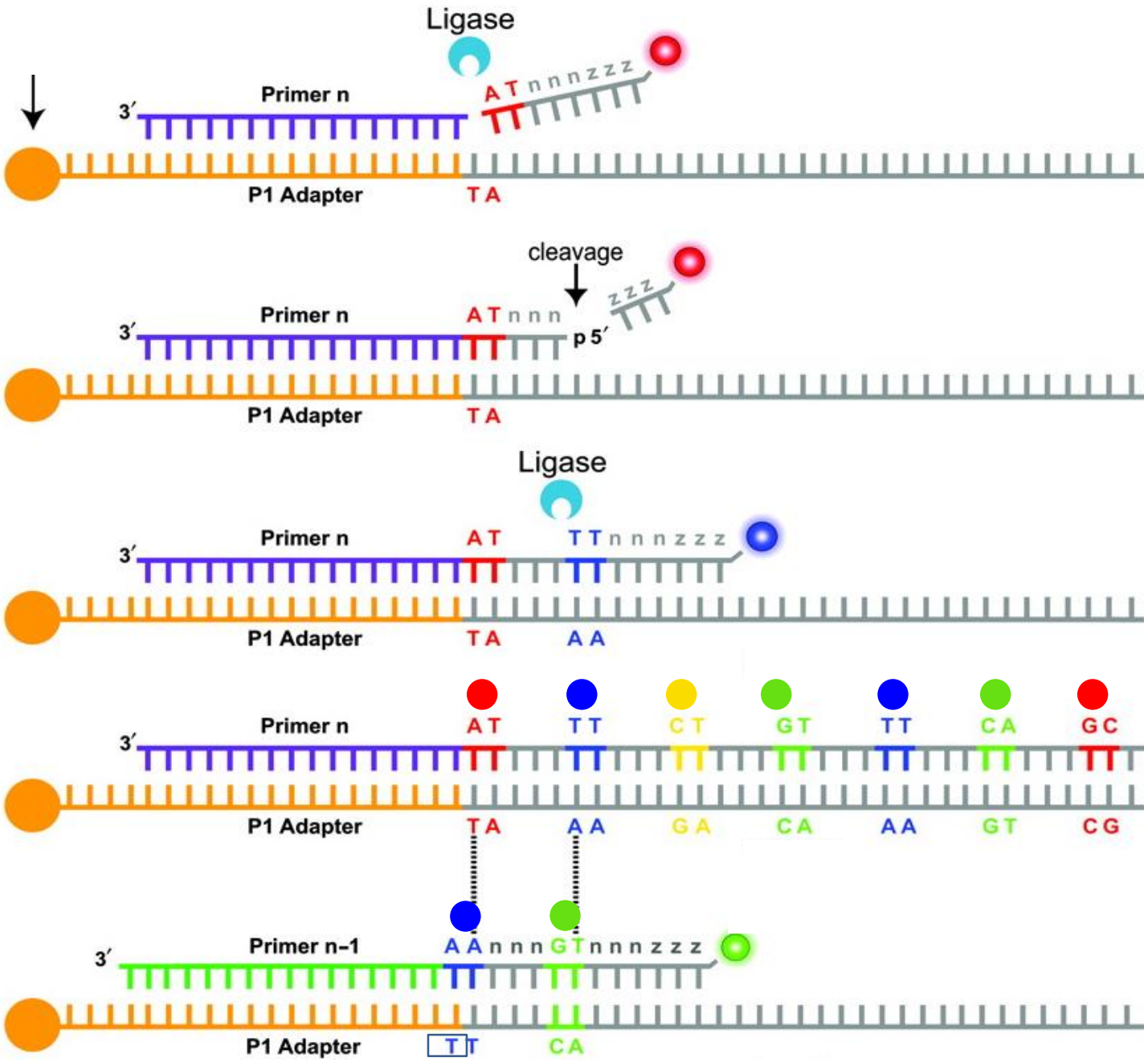
(Special attention to SOLiD Platform)




Primer n
3′
A T n n n z z z
P1 Adapter
T A
cleavage
A T n n n
z z z
p 5′
Ligase
T T n n n z z z
A A
A T
T T
C T
G T
T T
C A
G C
T A
A A
G A
C A
A A
G T
C G


3′

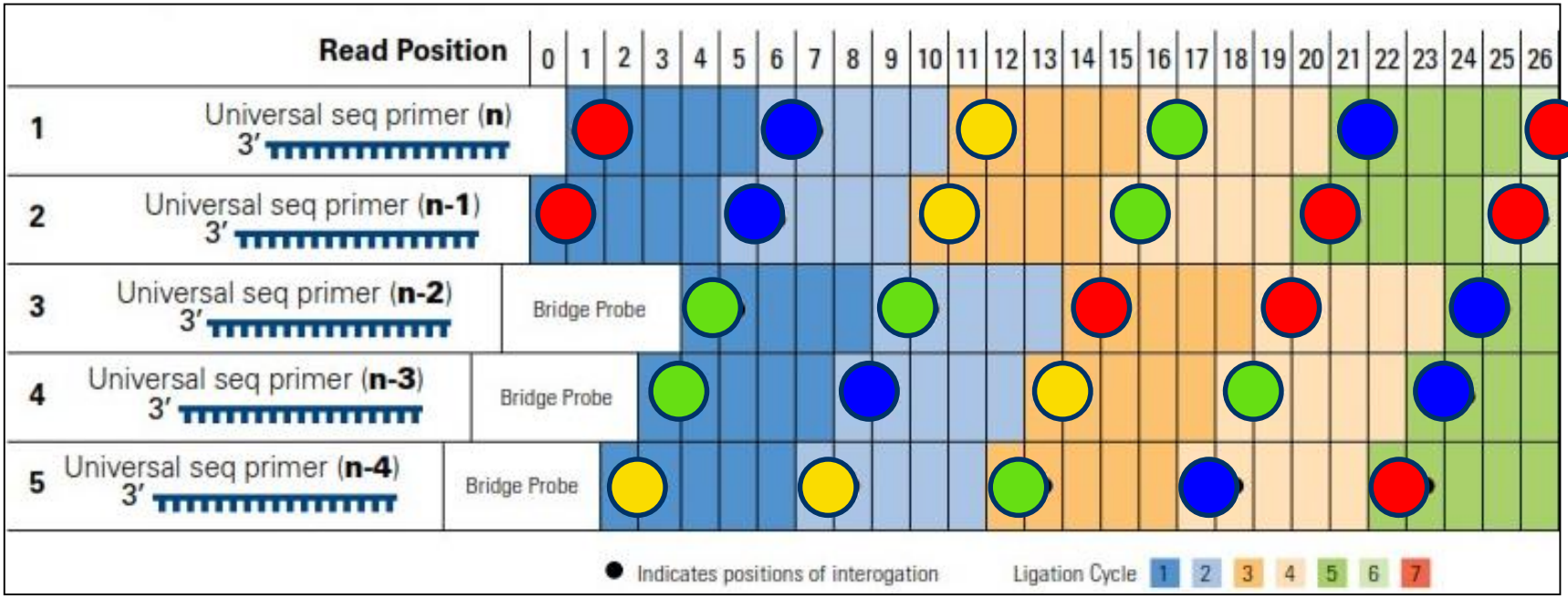

A C G T
A
C
G
T



known base as a part of the adapter

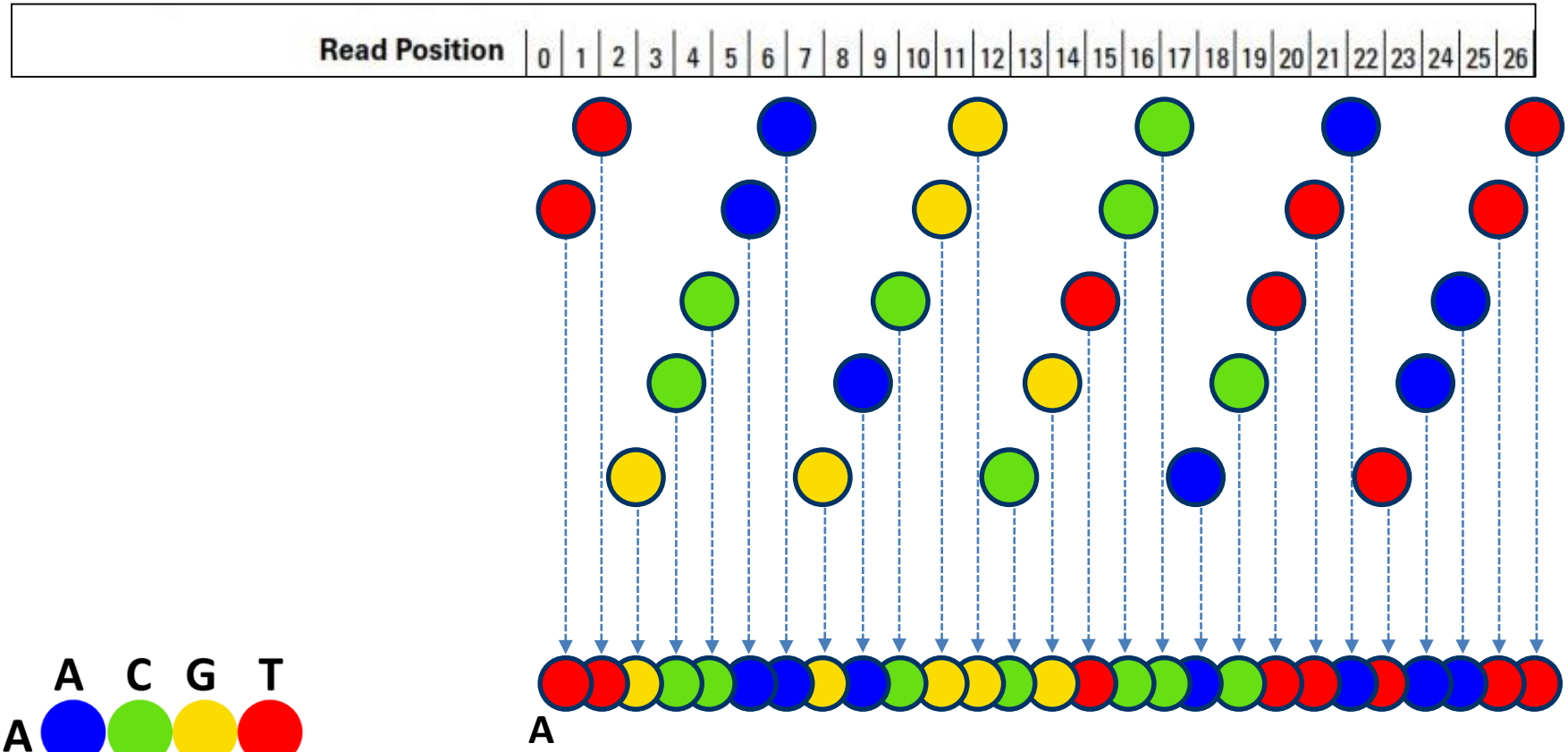
In NGS, for Base Calling

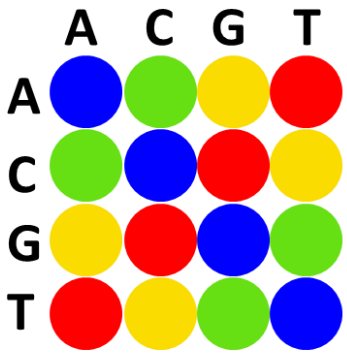
(Special attention to SOLiD Platform)



In NGS, for Base Calling

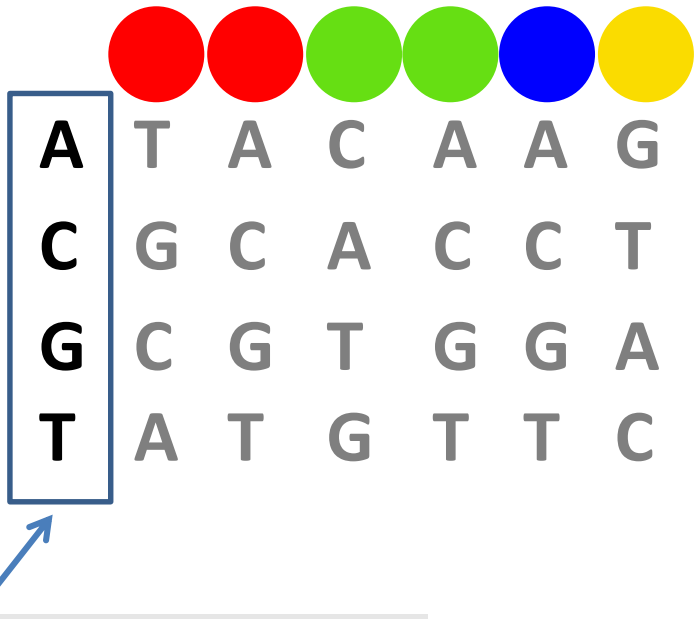
(Special attention to SOLiD Platform)





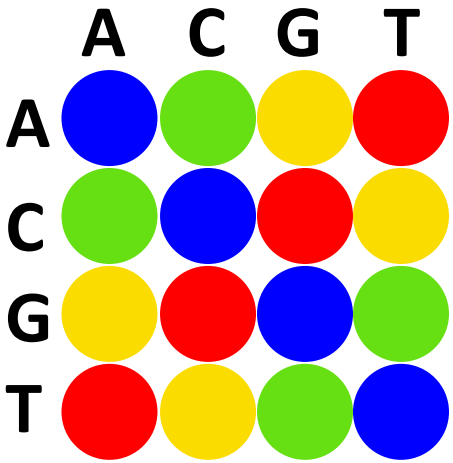
In NGS, for Base Calling

(Special attention to SOLiD Platform)

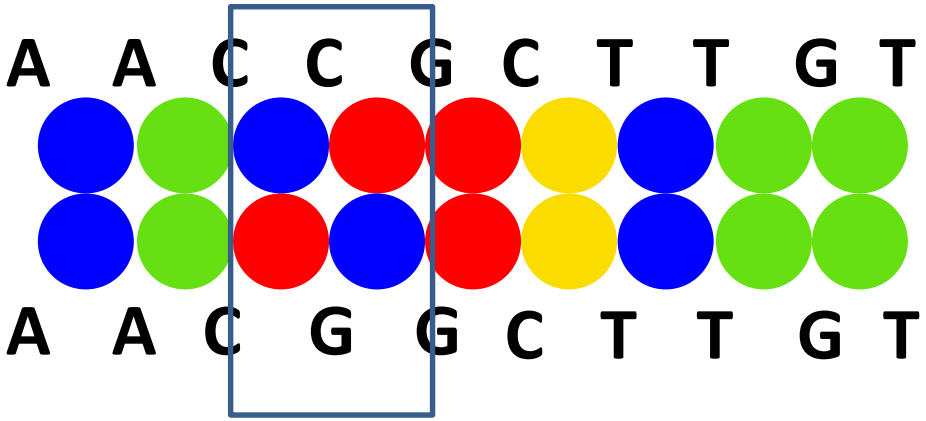


Must know identity of first base to decode color space

Strength in SNP



	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0



Reassurance of the mutation result with the adjacent color

Example

	A	C	G	T
A	0	1	2	3
C	1	0	3	2
G	2	3	0	1
T	3	2	1	0

T201101131310122310231112010003220 03

Summary of G-III

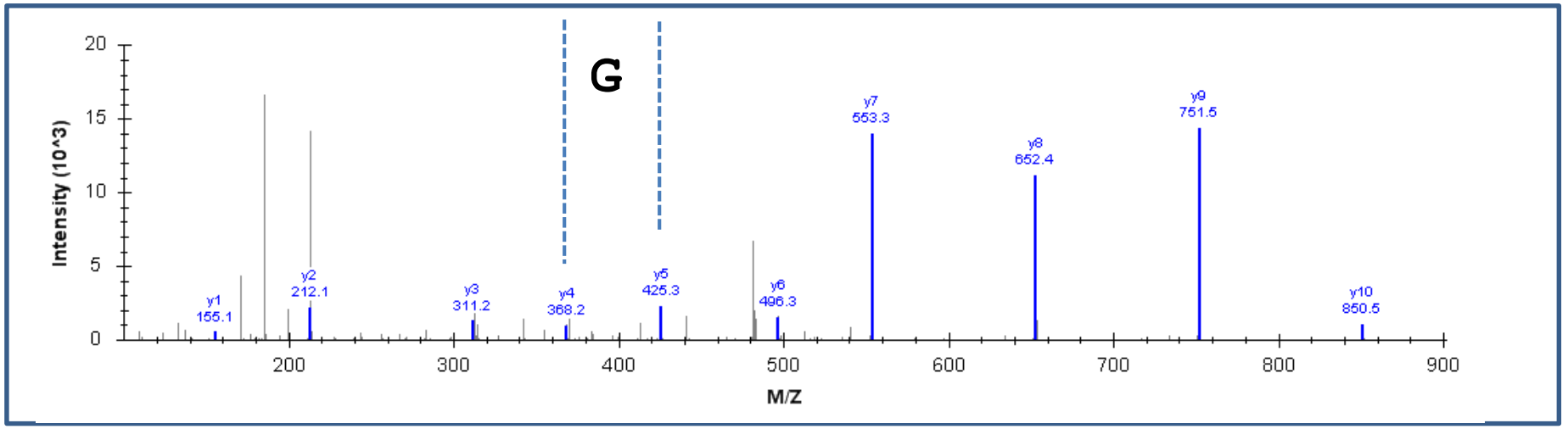
- Library construction: No needs of run e-coli cloning, isolation and picking.
  - Random fragmentation of starting DNA (sonication, nebulization or shearing) followed by DNA repair and end-polishing
  - Custom made, platform-specific adaptor ligation
  - Straightforward processing: quicker, cheaper
- Library amplification: Beads or glass (solid support)
- Sequencing and detection: No separation involved, direct step-by-step detection of each nucleotide base incorporated during the sequencing reaction.
- Throughput: Upto 100M reactions imaged per instrument run (Massive parallel sequencing)
- Shorter read lengths than capillary based sequencers
- Digital read type that enables direct quantitative comparison

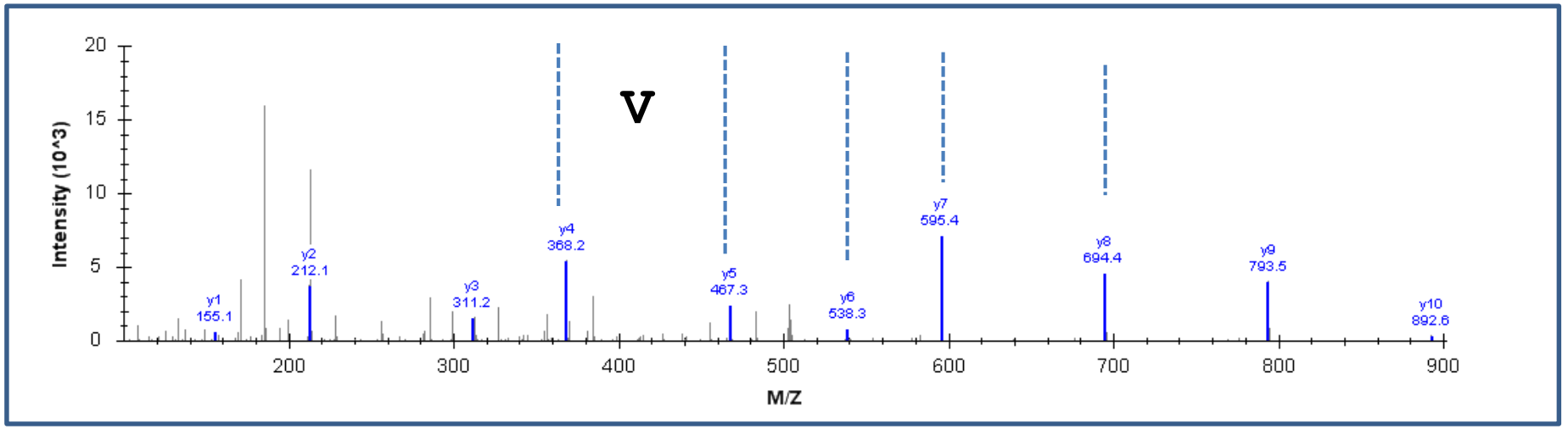
Slide adapted from Next-Generation Sequencing Technologies - Elaine Mardis (2012)

CASE III

LVVVGAGGVGK

LVVVGAVGVGK





nature biotechnology

A proteomics approach for the identification and cloning of monoclonal antibodies from serum

Wan Cheung Cheung[1,2], Sean A Beausoleil[1,2], Xiaowu Zhang[1], Shuji Sato[1], Sandra M Schieferl[1], James S Wieler[1], Jason G Beaudet[1], Ravi K Ramenani[1], Lana Popova[1], Michael J Comb[1], John Rush[1] & Roberto D Polakiewicz[1]

We describe a proteomics approach that identifies antigen-specific antibody sequences directly from circulating polyclonal antibodies in the serum of an immunized animal. The approach involves affinity purification of antibodies with high specific activity and then analyzing digested antibody fractions by nano-flow liquid chromatography coupled to tandem mass spectrometry. High-confidence peptide spectral matches of antibody variable regions are obtained by searching a reference database created by next-generation DNA sequencing of the B-cell immunoglobulin repertoire of the immunized animal. Finally, heavy and light chain sequences are paired and expressed as recombinant monoclonal antibodies. Using this technology, we isolated monoclonal antibodies for five antigens from the sera of immunized rabbits and mice. The antigen-specific activities of the monoclonal antibodies recapitulate or surpass those of the original affinity-purified polyclonal antibodies. This technology may aid the discovery and development of vaccines and antibody therapeutics, and help us gain a deeper understanding of the humoral response.

W. Cheung, Nature Biotech. ***30***, 447 (2012)

Discussion Point

1. Challenges of applying proteomics-based identification for Abs.
2. Reference database used in this study.
3. Why the used of 454 NGS was critical to this study (compared to CE-based sequencing?)