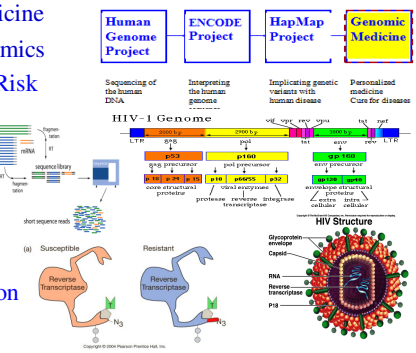
Computational Methods in Genomics
PART FOUR

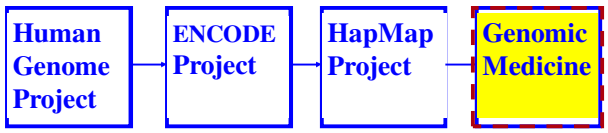
Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
khuri@cs.sjsu.edu
www.cs.sjsu.edu/faculty/khuri

Outline

- Genomic Medicine
- Pharmacogenomics
- DTC Disease Risk
- Biomarkers
- Retroviruses
- HIV (2008)
- Coreceptor
- CCR5Δ32
- Next Generation Sequencing



Pathway to Genomic Medicine



Sequencing of the human DNA	Interpreting the human genome sequence	Implicating genetic variants with human disease	Personalized medicine Cure for diseases

GWAS → Improved Health?

- Use of genetic information regarding common disease to individualize providers' approach to patients and change patients' behaviors in ways that lead to improved health ("Personalized Medicine").
- Use of genetic information regarding common disease to understand the biology of human disease to lead to improved diagnostic, therapeutic, and preventive approaches.

Personalized Medicine

Personalized medicine is the use of diagnostic and screening methods to better manage the individual patient's disease or predisposition toward a disease.

Personalized medicine will enable risk assessment, diagnosis, prevention, and therapy specifically tailored to the unique characteristics of the individual, thus enhancing the quality of life and public health.

Personalized Medicine is Genotype-Specific Treatment.

Variation in Medication Responsiveness

- Many human medications are not administered in their final and active form.
- The drugs are metabolized in a predictable way, and the enzymatic product is the therapeutic compound.
- People fall into one of 3 classifications:
  - Typical metabolizers
  - Poor metabolizers
  - Ultra-rapid metabolizers

Drug Studies and Dosages

- Drug studies are performed on large panels of people to determine the optimum dosage for the "average" person.
- However, any one person may not have the average metabolism, so the ideal dosage for him or her may not be the average dosage.
- When drugs are administered to different populations, it is important to determine a population-specific recommended dosage.

Cytochrome P450

- Cytochrome P450 is a family of enzymes (isozymes) that metabolize a large number of "pre-drugs".
- It is encoded by 2 separate genes:
  - 2D6: it has 9 exons and 8 introns, and is on chromosome 22
  - 2C19: it has 9 exons and 8 introns, and is on chromosome 10.

Cytochrome P450 2D6 (I)

- Twelve SNPs have been identified that lead to altered 2D6 protein activity.
  - The most common mutation is a G → A substitution within exon 4 that alters splicing in mRNA formation and results in no protein being produced.
- Over 40 pre-drugs require 2D6 protein activation, including heart medication, antidepressants, and painkillers.

Cytochrome P450 2D6 (II)

- Cytochrome P450 2D6 is involved in metabolizing painkilling medication such as codeine.
- 2-10% of the population are homozygous for null alleles and cannot use codeine for pain relief.
- It has been hypothesized that cytorochome P450 2D6 poor metabolizers are less tolerant of pain.

Cytochrome P450 2C19 (I)

- CYP2C19 (cytochrome P450 2C19) acts on 5-10% of drugs in current clinical use.
- About:
  - 2-6% of individuals of European origin (Caucasians),
  - 15-20% of Japanese, and
  - 10-20% of Africans

  have a slow acting, poor metabolizer form of this enzyme.

www.healthanddna.com/healthcare-professional

Cytochrome P450 2C19 (II)

- Cytochrome P450 2C19 (CYP2C19) is an isoenzyme of the cytochrome P450 super family and is responsible for the biotransformation (metabolism) and elimination of many commonly prescribed drugs including: anticonvulsants, antidepressants, cancer chemotherapy, antimalaria, antiulcer, and several proton pump inhibitors.
- Pharmacogenetic variation leads to inappropriate concentrations of drugs and drug metabolites, which may contribute to toxicity and risk of adverse drug reactions, or lack of therapeutic benefit.

www.aruplab.com/TestDirectory

Food and Drug Interactions

- Grapefruit juice can alter the ability of absorbing drugs.
- Pills we take pass through the stomach and dissolve in the small intestine, where the medication is absorbed.
- P-glycoprotein is a protein involved in pumping the drug we take into intestinal cells.
- Cytochrome P450 3A is a metabolizer which converts the drug into a more readily excreted form.

Blocking P-Glycoprotein & P450 A3

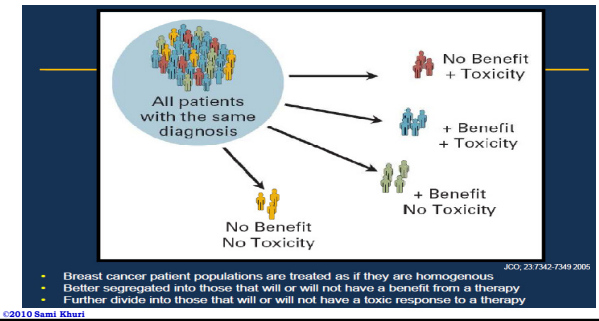
- One glass of grapefruit juice can block P-glycoprotein and cytochrome P450 3A for as long as 24 hours.
- Cytochrome P450 3A is inactivated by an unknown component of grapefruit juice, causing the enzyme to be destroyed.
- There is very little research in the area of genomic interactions with drugs and food, even though they affect human health.

Pharmacogenomics

- **Pharmacogenomics** deals with the influence of genetic variation on drug response in patients by correlating gene expression or SNPs with a drug's efficacy or toxicity.
- **Pharmacogenomics** aims to optimize drug therapy, with respect to the patients' genotype, to ensure maximum efficacy with minimal adverse effects.
- Such approaches promise the advent of "personalized medicine" in which drugs and drug combinations are optimized for each individual's unique genetic makeup.

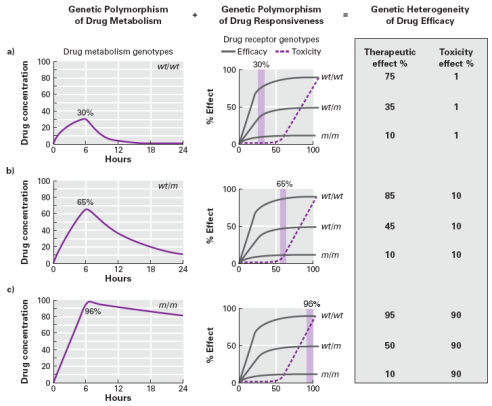
www.wikipedia.com

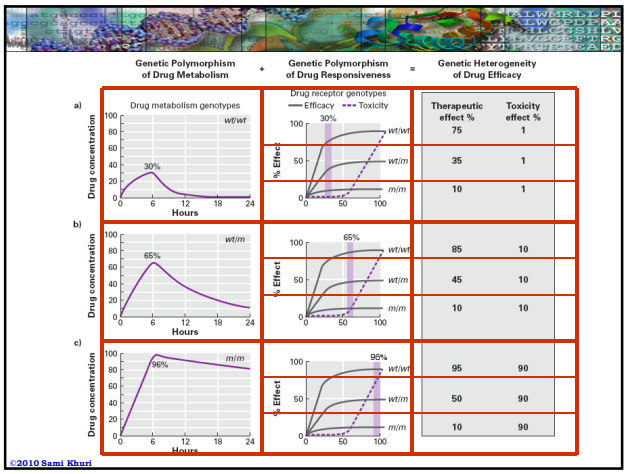
From Population to Subpopulation

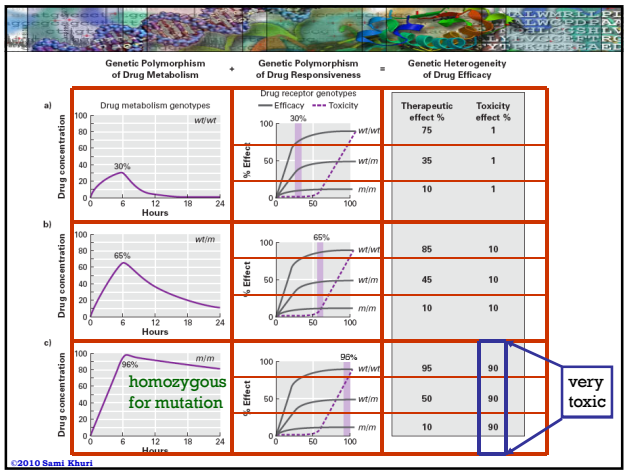


W. Evans and M. Relling

- Evans and Relling considered the efficacy and toxicity of a drug that requires two genes:
  - An activator with 2 alleles, and
  - A binding site with 2 alleles.
- There are 9 possible genotypes.
- Therapeutic effects depend on the genotype of the drug receptors in combination with the amount of active drug in circulation.







Therapeutic Effects of Drugs

- Therapeutic effects depend on the genotype of drug receptors in combination with the amount of active drug in circulation.
- The example highlights the complex web of protein interactions that pharmacogenomics hopes to decipher.
- Drug response is polygenic, and new technologies are needed to understand the connections between relevant proteins involved in drug responses.

©2010 Sami Khuri

Clinically Relevant SNPs

- Traditionally, drug development has been aimed at delivering medications that are effective and safe for everyone.
  - But enzyme polymorphism can have clinically significant consequences.
- Pharmaceutical companies are spending a lot of money to discover clinically relevant SNPs in order to produce SNP haplotype-specific medications.

©2010 Sami Khuri

Genotype-Specific Medication

- If genotype-specific medication becomes viable, when a person is diagnosed with an illness, the physician will need to know the genotype of the person to determine the appropriate medication and dosage for optimal therapy.
- Pharmacagenomics is not as futuristic as it may sound as we see in the Iressa case.

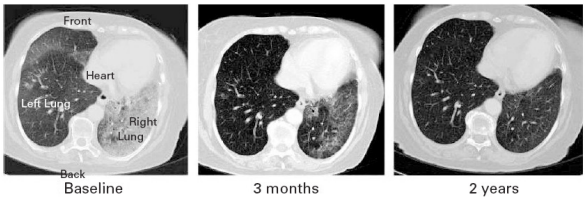
©2010 Sami Khuri

Non-Small-Cell Lung Cancer

- Every year 140,000 patients are diagnosed with non-small-cell lung cancer, which is nearly always fatal.
- During clinical trials of Iressa, about 10% of patients were completely cured, but all others died.
- A mutation in the epidermal growth factor receptor (EGFR) gene determines if Iressa will cure or not.

©2010 Sami Khuri

Iressa: To Cure or Not to Cure



Three CAT scans of a person who is cured by taking Iressa:
The right lung is hazy with invasive non-small-cell lung cancer.
Within 3 months, the cancerous lung is clearing.
Two years later, the cancer is gone.

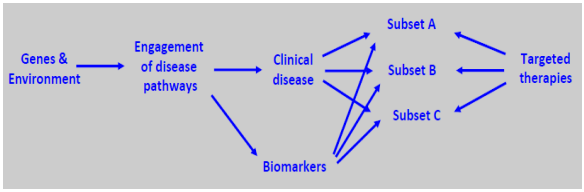
Some Diseases Involve Many Genes

- There are a number of classic "genetic diseases" caused by mutations of a single gene
  - Huntington's, Cystic Fibrosis, Tay-Sachs, PKU, etc.
- There are also many diseases that are the result of the interactions of many genes:
  - asthma, heart disease, cancer
- Each of these genes may be considered to be a **risk factor** for the disease.
- Groups of genetic markers (SNPs) may be associated with a disease without determining a mechanism.

SNPs as Biomarkers

- A lot of effort has been focused on discovering SNPs that are in the proximity of genes.
- The hope is that identifying such SNPs will lead to the diagnosis and treatment of more diseases more effectively.
- However, this task is rendered more problematic by the realization that drug effectiveness is hampered by genomic variations.

Biomarkers and Human Disease

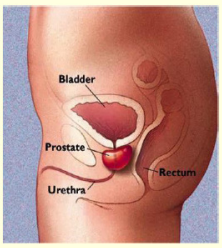


- Improve clinical trial design
- Identify disease subsets
- Guide disease selection for new therapies

Prostate Cancer Diagnosis

Risk factors 2004

- Age
- Race
- Family history



Risk factors 2008

- Age
- Race
- Family history
- Locus 1
- Locus 2
- .....
- .....
- Locus 16

cancercontrol.cancer.gov/od/phg/presentations/Xu.pdf

GWAS and Prostate Cancer

Prostate cancer risk associated variants identified from GWAS

SNPs	Chr	Position	Allele frequency Cases	Allele frequency Controls	OR (95% CI)	P
rs2660753	3p12	87,193,364	0.10	0.08	1.32 (1.13-1.54)	3.4E-04
rs9364554	6q25	160,804,075	0.33	0.31	1.12 (1.02-1.22)	0.02
rs10486567	7p15	27,749,803	0.78	0.76	1.12 (1.01-1.24)	0.03
rs6465657	7q21	97,654,263	0.51	0.47	1.16 (1.06-1.26)	6.7E-04
rs16901979	8q24 (2)	128,194,098	0.06	0.03	1.66 (1.34-2.07)	3.1E-06
rs6983267	8q24 (3)	128,482,487	0.56	0.51	1.22 (1.12-1.33)	3.6E-06
rs1447295	8q24 (1)	128,554,220	0.17	0.14	1.21 (1.08-1.36)	1.6E-03
rs10993994	10q11	51,219,502	0.43	0.39	1.15 (1.05-1.25)	1.6E-03
rs10896449	11q13	68,751,243	0.49	0.46	1.14 (1.05-1.25)	2.1E-03
rs4430796	17q12	33,172,153	0.61	0.56	1.24 (1.14-1.35)	8.5E-07
rs1859962	17q24.3	66,620,348	0.54	0.50	1.17 (1.08-1.28)	2.0E-04
rs5945619	Xp11	51,074,708	0.42	0.38	1.20 (1.06-1.36)	3.5E-03

cancercontrol.cancer.gov/od/phg/presentations/Xu.pdf

Genetic Testing for the Public

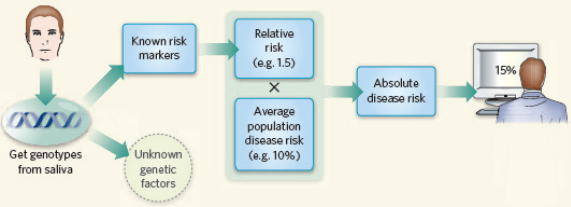
23andMe

Navigenics™

deCODEme

23andme.com
decodeme.com
navigenics.com

Direct-to-Consumer Disease Risk



- More than 1,000 DNA variants associated with diseases and traits have been identified.
- Direct-to-consumer (DTC) companies are harnessing these discoveries by offering DNA tests that provide insights

"An agenda for personalized medicine" by P. Ng et al., Nature, October 2009

Pharmaceutical Companies

- Most of the major pharmaceutical companies are currently collecting pharmacogenomic data in their clinical trials.
- Data is yet to be published.
- Genetic indications for drug use are still a couple of years away.
- Plan to sell the drug with the gene test

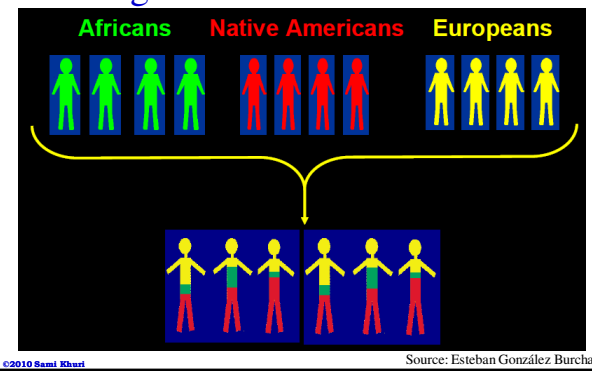
Genotyping Costs

	Number of SNPs	Cost per sample*
Illumina HumanHap 1M	1,000,000	$580-650
Illumina HumanHap 610K	610,000	$480-520
Illumina HumanHap 550K	550,000	$290-370
Affymetrix 6.0	1,000,000	$~400
Affymetrix 5.0	500,000	$~300
Illumina iSelect	28,000	$90-230
Illumina iSelect	7,600	$65-110
Illumina GoldenGate (OPA)	1,536	$85-200
Fluidigm/BioMark	48	$10+
TaqMan	1	$0.60

*approximate list prices (02/14/2008) for reagents only.

GWAS Bioinformatics by Kevin Jacobs

Origins of African Americans



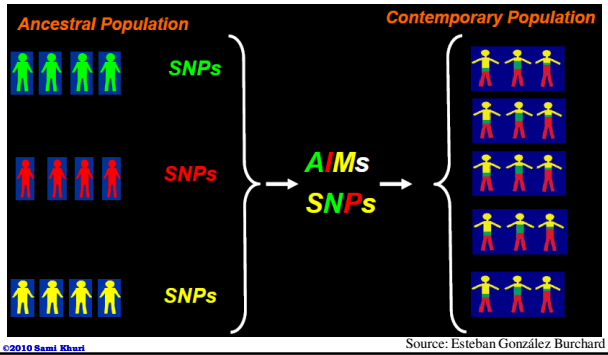
Source: Esteban González Burchard

Ancestry Informative Marker

- An **Ancestry-Informative Marker** (AIM) is a set of polymorphisms for a locus which exhibits substantially different frequencies between populations from different geographical regions.
- By using a number of **AIM**s one can estimate the geographical origins of the ancestors of an individual and ascertain what proportion of ancestry is derived from each geographical region.

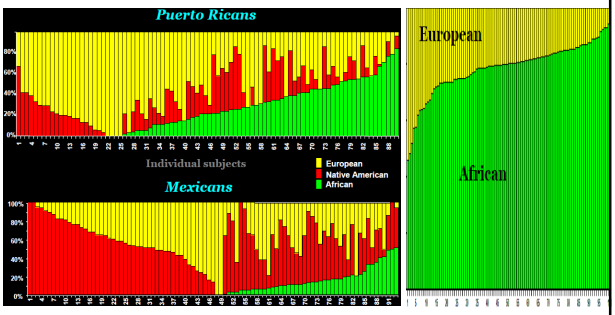
en.wikipedia.org/wiki/

SNPs and AIMs



Source: Esteban González Burchard

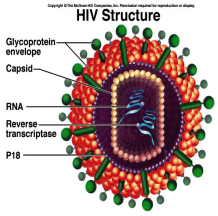
Origins of Latinos and African Amer.



Source: Esteban González Burchard

Self-Identified Race: Genetic Ancestry

Retroviruses



- Intracellular Parasites
- gag, pol, & env genes
- HIV-1
- Immune System
- Macrophage
- T-Cells
- AIDS

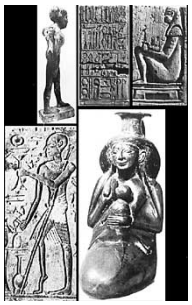
What is a Virus?

Viruses: Small living particles that can infect cells and change how the cells function. Infection with a virus can cause a person to develop symptoms.

The disease and symptoms that are caused depend on the type of virus and the type of cells that are infected.

www.medterms.com

Ancient Egypt and Viruses

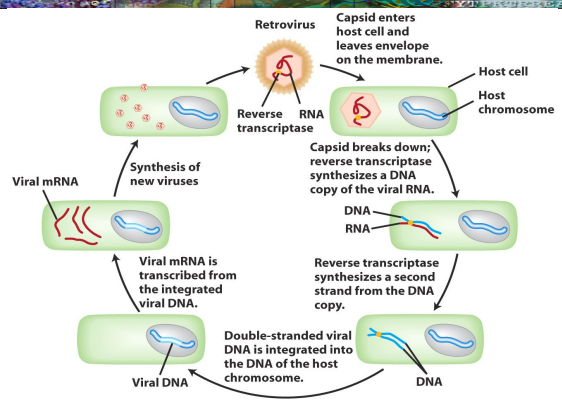


HIV Case Study

- Why have promising AIDS treatments, like drug azidothymidine (AZT) proven ineffective in the long run?
- Why does HIV kill people?
- Why are some people resistant to becoming infected or to progressing to disease once they are infected?
- Where did HIV come from?

Retrovirus

- A **retrovirus** is a single-stranded RNA virus that employs a double-stranded DNA (dsDNA) intermediate for replication.
- The RNA is copied into DNA by the enzyme **reverse transcriptase**.
- The dsDNA is integrated into the host chromosomes, from which it is transcribed to produce the viral genome and proteins that form new viral particles.



HIV

- The human immunodeficiency virus (**HIV**) is the virus that causes acquired immune deficiency syndrome (**AIDS**).
- **HIV** moves from person to person when a bodily fluid containing the virus, usually blood or semen, carries the virus from an infected person directly onto a mucous membrane or into the bloodstream of an uninfected person.

What is HIV?

- Like all viruses, **HIV** is an intracellular parasite.
- It is incapable of an independent life and is highly specific in the cell types it afflicts.
- **HIV** parasitizes components of the human immune system: macrophages and T cells.
- **HIV** uses the enzymatic machinery and energy found in these cells to make copies of itself, killing the host cells in the process.

http://www.niaid.nih.gov/factsheets/howhiv.htm

Macrophages and T Cells

- **Macrophage** - a large immune system cell that devours invading pathogens and other intruders. Stimulates other immune system cells by presenting them with small pieces of the invaders.
- **CD4+ T cells** - white blood cells that orchestrate the immune response, signaling other cells in the immune system to perform their special functions. Also known as T helper cells, these cells are killed or disabled during HIV infection.

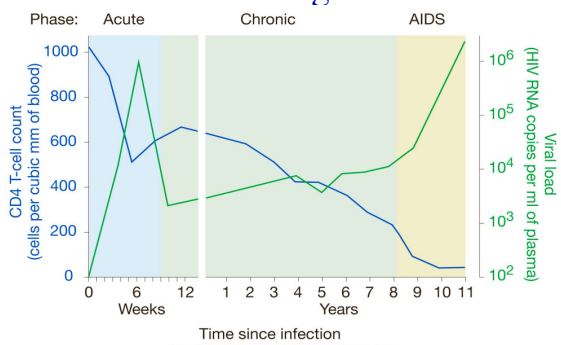
HIV is a Lentivirus

- HIV is a retrovirus that belongs to the class of lentiviruses:
  - Lentiviruses are slow viruses. The course of infection with these viruses is characterized by a long interval between initial infection and the onset of serious symptoms.
- Other lentiviruses infect nonhuman species.
  - Example
    - Feline immunodeficiency virus (**FIV**) infects cats
    - Simian immunodeficiency virus (**SIV**) infects monkeys and other nonhuman primates.

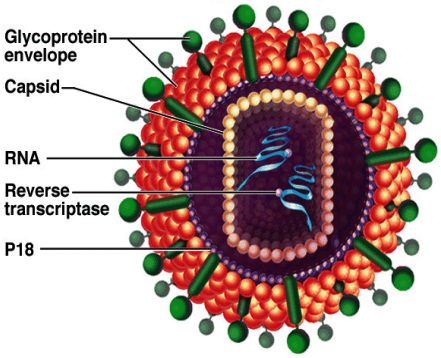
How Does HIV Cause AIDS?

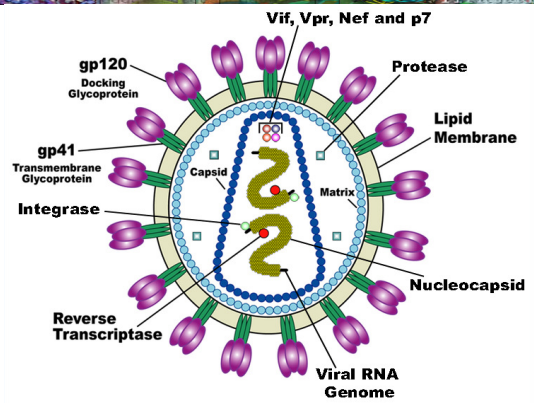
- The human body responds to HIV infection by destroying virions floating in the bloodstream and by killing its own infected cells before new virions are assembled and released.
- Ultimately, the supply of CD4 helper T cells depletes and the immune system collapses.

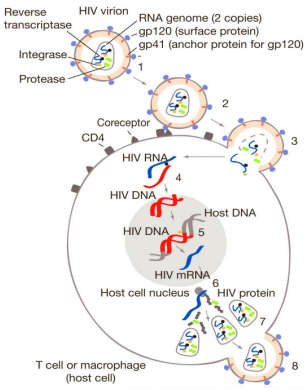
Disease Progression



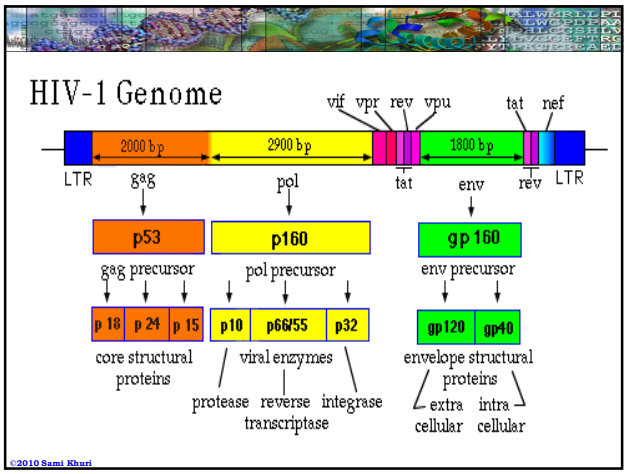
HIV Structure







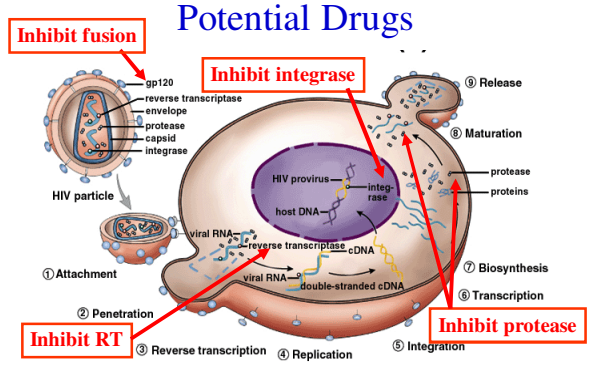
1) HIV's extracellular, or virion stage
2) HIV's gp120 protein binds to CD4 and coreceptor on host cell
3) HIV's RNA genome, reverse transcriptase, integrase, and protease enter host cell
4) Reverse transcriptase synthesizes HIV DNA from HIV's RNA template
5) Integrase splices HIV DNA into host genome. HIV DNA is transcribed to HIV mRNA by the host cell's RNA polymerase
6) HIV mRNA is translated to HIV precursor proteins by host cell's ribosomes. Protease cleaves precursors into mature viral proteins
7) New generation of virions assembles inside host cell
8) New virions bud from host cell's membrane



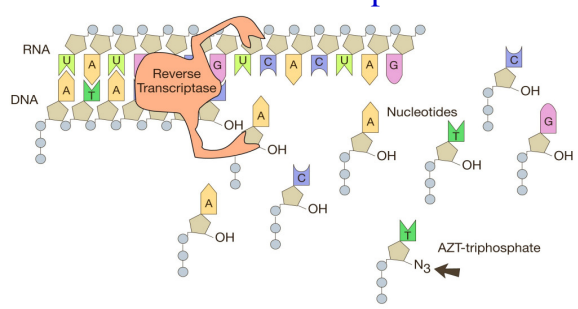
Difficulty of HIV Treatment

- **HIV** uses the host cell's own enzymatic machinery:
  - its polymerase
  - its ribosome
  - its tRNAs
- Drugs that interrupt the life cycle of the virus are almost certain to interfere with the host's cell enzymatic as well and thus causing serious side effects.

Potential Drugs



Azidothymidine (AZT) Blocks Reverse Transcriptase

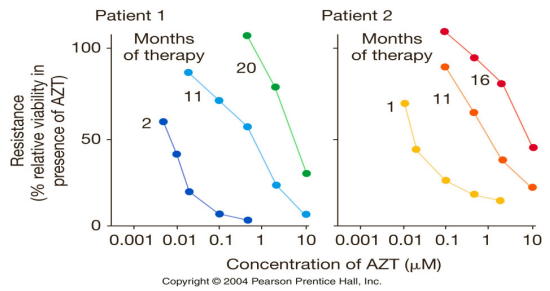


Azidothymidine Results

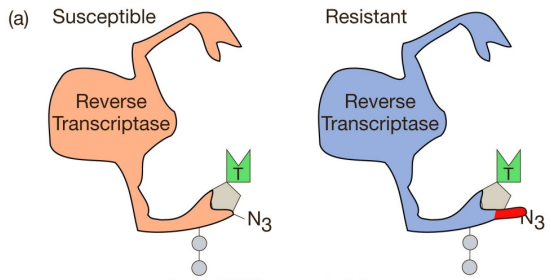
- AZT worked in early tests:
  - Effectively halted the loss of macrophages and T cells in AIDS patients.
- AZT can cause serious side effects because it sometimes fools DNA polymerase and interrupts DNA synthesis in host cells.
- After a few years of use, patients stop responding to treatment.

AZT Resistance

(a) Resistance of HIV in two patients, followed over time



Mutations in Reverse Transcriptase



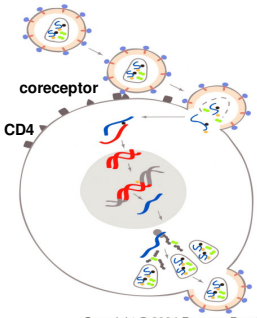
Some People are Resistant to HIV

- In the early 1990s, work from several laboratories demonstrated that some people remain uninfected even after repeated exposure to the virus and some people who are infected with the virus survive many years longer than expected.
- Resistant individuals have unusual forms of the coreceptor molecules and these mutant proteins thwart HIV entry.

Human CC-CKR-5

- CC-CKR-5 gene is located on chromosome 3
- CC-CKR-5 gene encodes a protein called C-C chemokine receptor-5, abbreviated CCR5
  - CCR5 is a cell surface protein found on white blood cells.
  - The CCR5 function is to bind chemokines, which are molecules released as signals by other immune system cells.
    - When a white blood cell is simulated by chemokines binding to its receptors, the cell moves into inflamed tissues to help fight an infection.
  - CCR5 is also exploited as a coreceptor by most sexually transmitted strains of HIV-1

CCR5 Function in HIV-1 Infection



HIV entry into the cell requires binding to a CD4 molecule and, in the majority of cases, to a coreceptor, either chemokine coreceptor 4 (CXCR4) or 5 (CCR5).

Rong Liu et al.

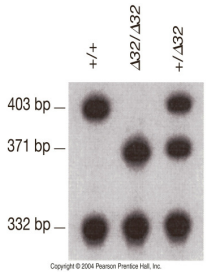
- A CKR-5 allele present in the human population appears to protect homozygous individuals from sexual transmission of HIV-1 strain R5.
- These individuals appear to have inherited a defective CKR-5 allele that contains an internal 32 base pair deletion.
- The deletion occurs within the coding region and results in a frame shift.
- The encoded protein is severely truncated and cannot be detected at the cell surface.

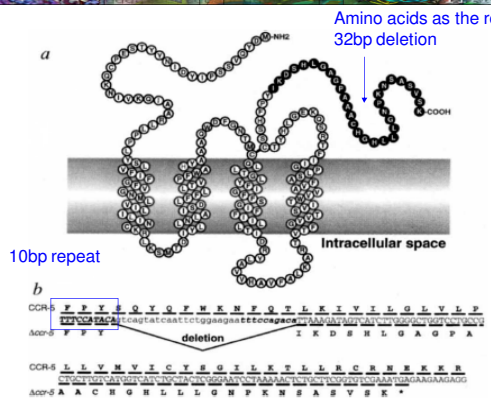
Determining CCR5 Genotypes

- Functional allele is CCR5+, or just +
- The allele with 32-bp deletion is CCR5-Δ32, or just Δ32
- Individuals with +/+ genotype are susceptible to HIV-1
- Individuals with +/Δ32 genotype are susceptible, but may progress to AIDS more slowly
- Individuals with Δ32/Δ32 genotype are resistant to HIV-1 R5

Genotyping Individuals

- DNA gel electrophoresis of the pattern resulting from PCR amplification and EcoRI cleavage.
- A 735-bp PCR product is cleaved into a common band of 332 for both alleles and into 403-bp and 371-bp bands for the + allele and Δ32 alleles, respectively.





Genotyping Individuals

- Samson et al. took DNA samples from a large number of individuals from different parts of the world, examined the gene for CCR5 in each individual and calculated the frequency of the normal and Δ32 alleles in each population.

Calculating Allele Frequencies

- For example, to calculate the frequency of the Δ32 allele in the Ashkenazi population in Europe from Martinson et al. (1997):
- 43 individuals were tested:
  - 26 were homozygous for + allele
  - 1 was homozygous for Δ32 allele
  - 16 were heterozygous
- Genotype frequencies are:

+/+:	26/43 = 0.605
+/Δ32:	16/43 = 0.372
Δ32/Δ32:	1/43 = 0.023

Calculating Allele Frequencies

- Genotype frequencies are:

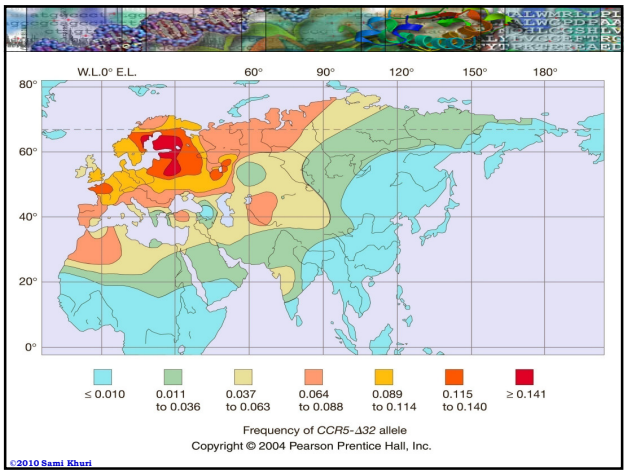
+/+:	26/43 = 0.605
+/Δ32:	16/43 = 0.372
Δ32/Δ32:	1/43 = 0.023

- The frequency of the Δ32 allele is the frequency of Δ32/Δ32 plus half the frequency of +/ Δ32:

  0.023 + ½ * 0.372 = 0.209

CCR5-Δ32 Allele Distribution

- Gene frequency of about 10% was observed for CCR5-Δ32 in populations of European descent.
- As we move away from northern Europe, both to the east and to the south, the frequency of the Δ32 allele declines.
- Outside of Europe, Middle East, and western Asia, the Δ32 allele is virtually absent.



Why Two Forms of CCR5?

- Why would one form of a gene be relatively common in one population, but absent in others?
- Two possible explanations:
  - The CCR5-Δ32 allele may have been recently favored by natural selection in European populations; or
  - The allele could have risen to high frequency by chance, in a process called genetic drift.

Natural Selection Hypothesis

- The Δ32 allele confers protection against a pathogen other than HIV, such as bubonic plague or smallpox.
- The Δ32 allele would have risen to high frequency because of the survival advantage it offered during devastating epidemics that swept Europe during the past millennium.

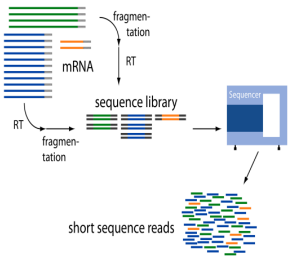
Genetic Drift Hypothesis

- The Δ32 allele first appeared and achieved a high frequency among the Vikings and then was disseminated across Europe during the Vikings raids of the 8th, 9th, and 10th centuries.

Coreceptor Antagonists

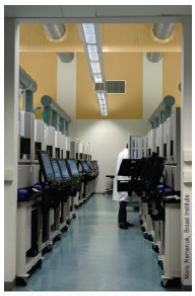
- Molecular biologists are trying to design drugs that mimic the effect of the resistance alleles.
- One approach is to find small molecules that bind to the CCR5 protein on the surface of host cells and block HIV's attempt to use the protein as coreceptor:
  - Maraviroc is the first CCR5 coreceptor antagonist to receive marketing approval from the Food and Drug Administration (FDA) for the treatment of CCR5-tropic human immunodeficiency virus (HIV) infection as part of an optimized antiretroviral regimen in treatment-experienced patients.

Next Generation Sequencing Techniques

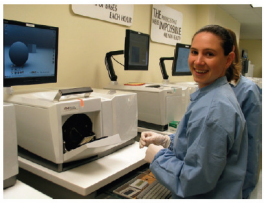


- Roche 454
- Illumina Solexa
- AB SOLiD
- Emulsion
- Pyrosequencing
- Paired Reads
- Transcriptome

Next Generation Sequencing Tech.



Sequencing centers producing the Sanger sequence data for mammalian genome projects are factory-like outfits with a large number of personnel.



The latest next-generation sequencing instruments can generate as much data in 24 h as several hundred Sanger-type DNA capillary sequencers, but are operated by a single person.

Stephan Shuster 2008

Bioinformatics and NGS

- NGS technologies are revolutionizing the scale and perspectives of research in the fields of genomics and functional genomics.
- The general features of the three major NGS platforms, namely Roche 454, Illumina Solexa and AB SOLID, are illustrated.
- NGS data require 'next-generation bioinformatics' for the handling and the analysis of the huge amount of data produced.
- A simulation carried out by using two benchmarks datasets against the human genome and transcriptome illustrates current limitations and open problems in genome mapping of NGS data.
- The major bioinformatics applications for dealing with NGS including genome mapping, *de novo* assembly, detection of SNPs and editing sites, transcriptome analysis, ChIP-Seq, small RNA characterization and epigenomic studies are briefly discussed.

Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing by Horner et al., 2009

Current Trends

- Next generation sequencing (NGS) techniques have been proposed:
  - high-throughput sequencing
  - massively parallel sequencing
  - flow-cell sequencing
- Sequencing devices are commercially available from:
  - Roche (formerly: 454)
  - Illumina (formerly: Solexa) of San Diego, CA: "GenomeAnalyzer"
  - Applied Biosystems (ABI) of Carlsbad, CA: "SOLiD system"
  - Helicos of Cambridge, MA: "Helicoscope"

Roche 454

- Presented in 2005
- emulsion PCR
- pyrosequencing (polymerase-based)
- read length: 250 bp
- paired read separation: 3 kb
- 300 Mb per day
- $60 per Mb
- error rate: around 5% per bp
- dominant type of error: indels

Illumina

- Second on the market
- bridge PCR
- polymerase-based sequencing-by-synthesis
- 32-40 bp (newest models: up to 100 bp)
- paired read separation: 200 bp
- 400 Mb per day (getting better)
- $2 per Mb
- error rate: 1% per bp (good reads: 0.1%)
- dominant error type: substitutions

Applied Biosystems SOLiD

- Since late 2007
- emulsion PCR
- ligase-based sequencing
- read length: 50bp
- paired read separation: 3 kb
- 600 Mb per day
- $1 per Mb
- very low error rate: <0.1% per bp (still high compared to Sanger capillary sequencing: 0.001%)
- dominant error type: substitutions (due to color shift)

Helicos ("Helicoscope")

- On the market since a 2007
- no amplification
- single-molecule polymerase-based sequencing
- read length: 25-45 bp
- 1200 Mb per day
- $1 per Mb
- error rate: <1% (manufacturer claim)

Polonator

- On the market since less than a year
- emulsion PCR
- ligase-base sequencing
- very short read-length: 13 bp
- but: low-cost instrument ($150,000)
- <$1 per Mb

Nextgen vs. Sanger Sequencing

- Two main differences between next generation and Sanger capillary sequencing:
  - The library is not constructed by cloning, but by a novel way of doing PCR, where the fragments are separated by physico-chemical means (emulsion PCR or bridge PCR).
- Many fragments are sequenced in parallel in a flow cell (as opposed to a capillary), observed by a microscope with Charge Coupled Device (CCD) camera.

Uses for Nextgen Sequencing

- De-novo sequencing and assembly of small genomes
- Transcriptome analysis (RNA-Seq, sRNA-Seq, ...)
  - Identifying transcribed regions
  - Expression profiling
- Resequencing to find genetic polymorphisms:
  - SNPs, micro-indels
  - CNVs
- ChIP-Seq, nucleosome positions, etc.
- DNA methylation studies (after bisulfite treatment)
- Environmental sampling (metagenomics)

RNA-Seq and ChIP-Seq

- RNA-Seq:
  - processed mRNA is converted to cDNA and sequenced,
  - is enabling the identification of previously unknown genes and alternative splice variants
- ChIP-Seq:
  - sequences immunoprecipitated DNA fragments bound to proteins,
  - is revealing networks of interactions between transcription factors and DNA regulatory elements
- The whole-genome sequencing of tumor cells is uncovering previously unidentified cancer-initiating mutations

Paired-end Sequencing

- The two ends of the fragments get different adapters.
- Hence, one can sequence from one end with one primer, then repeat to get the other end with the other primer.
- This yields "pairs" of reads, separated by a known distance (200bp for Illumina).

Uses of Paired-end Sequencing

- Paired-end sequencing is useful:
  - to find micro-indels
  - to find copy-number variations
  - to look for splice variants

Need for Bioinformatics

- New generation DNA sequencers provide billions of bases rapidly and inexpensively:
  - Illumina/Solexa: 75-75bp read pairs, 100 million in a run
  - ABI/SOLiD: similar in scale (50-50bp)
  - Roche/454: ~300-500bp reads, 100Mbp a run
- New algorithms are required for:
  - Alignment (read mapping)
  - Assembly
  - Statistical tests
  - Visualization

'Mapping' the Reads

- In contrast to whole-genome assembly, in which these reads are assembled together to reconstruct a previously unknown genome, many of the next-generation sequencing projects begin with a known, or so-called 'reference', genome.
- To make sense of the reads, their positions within the reference sequence must be determined.
  - This process is known as aligning or 'mapping' the read to the reference.

Read Mapping Problems

- In one version of the mapping problem, *short-read mapping problem*, reads must be aligned without allowing large gaps in the alignment.
- A more difficult version of the problem, *spliced-read mapping problem*, arises primarily in RNA-Seq, in which alignments are allowed to have large gaps corresponding to introns.

Challenges of Mapping Short Reads

- Need very efficient algorithms, in which every bit of memory is used optimally or near optimally.
  - if the reference genome is very large, and if we have billions of reads, how quickly can we align the reads to the genome?
    - DNA sequencers produce millions of reads per run.
    - Complete assays may involve many runs.
  - The recent cancer genome sequencing project by Ley *et al.* generated nearly 8 billion reads from 132 sequencing runs.
    - A large, expensive computer grid might map the reads from this experiment in a few days.

Mapping Efficiency

Program	BAC on MHC-162k	BAC on chr6	BAC on all
BLAST	06:56:11 (51M)	>5 days	>8 days
BLAT	00:04:06 (32M)	06:33:03 (32M)	7 days+22:47:16(32M)
RMAP	00:00:51 (1.9G)	00:27:54 (1.9G)	10:09:03 (1.9G)
Mosaik	00:05:33 (214M)	00:07:41 (3.4G)	02:11:15 (3.5G)
ZOOM	00:00:37 (1.1G)	00:06:09 (1.1G)	01:33:03 (1.1G)

Time is represented as hh:mm:ss.

BAC dataset: 3 415 291 reads; Lin, H. *et al.*, 2008

Repeat Challenges

- Need a strategy for resolving repeats.
  - if a read comes from a repetitive element in the reference, a program must pick which copy of the repeat the read belongs to.
    - The program must report multiple possible locations for each read or to pick a location heuristically.
  - Sequencing errors or variations between the sequenced chromosomes and the reference genome exacerbate this problem:
    - the alignment between the read and its true source in the genome may actually have more differences than the alignment between the read and some other copy of the repeat.

Additional Challenges

- Read errors:
  - dominant cause for mismatches in the alignment
  - detection of substitutions?
  - Importance of the base-call quality ("phred scores")
- Unknown reference genome
  - "de-novo" assembly

Short Read Mappers

- In the last few years, many tools for short-read alignments have been published:

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

Trapnell, C. & Salzberg, S.L., 2009

Indexing is the Key Strategy

- Short read mappers use a computational strategy known as 'indexing' to speed up their mapping algorithms.
- An index of a large DNA sequence allows one to rapidly find shorter sequences embedded within it.

Maq

- Maq is based on a straightforward but effective strategy called spaced seed indexing.
- Each read is divided into four segments of equal length, called the 'seeds'.
  - If the entire read aligns perfectly to the reference genome, then clearly all of the seeds will also align perfectly.
  - If there is one mismatch, however, perhaps due to a single-nucleotide polymorphism (SNP), then it must fall within one of the four seeds, but the other three will still match perfectly.
  - Two mismatches will fall in at most two seeds, leaving the other two to match perfectly.

Aligning Seed Pairs in Maq

- By aligning all possible pairs of seeds (six possible pairs) against the reference, it is possible to determine the list of candidate locations within the reference, where the full read may map, allowing at most two mismatches.
- The resulting set of candidate reads is typically small enough that the rest of the read—that is, the other two seeds that might contain the mismatches—may be individually checked against the reference.

Differences Between Tools

- Alignment tools differ in:
  - Speed
  - suitability for use on compute clusters
  - memory requirements
  - Accuracy:
    - Is a good match always found?
    - What is the maximum number of allowed mismatches?
  - ease of use
  - available down-stream analysis tools
    - Are there SNP and indel callers that can deal with the tool's output format?
    - Is there an R package to read in their output?

Additional Differences

- Alignment tools also differs in whether they can:
  - make use of base-call quality scores
  - estimate alignment quality
  - work with paired-end data
  - report multiple matches
  - work with longer than normal reads
  - match in color space (for SOLiD systems)
  - align data from methylation experiments
  - deal with splice junctions

Short-read Alignment Ideas

- Short-read alignment algorithms use one of these ideas:
  - use spaced seed indexing
    - hash seed words from the reference
    - hash seed words from the reads
  - sort reference words and reads lexicographically
  - use the Burrows-Wheeler transform (BWT)
  - use the Aho-Corasick algorithm

BWT

- The Burrows-Wheeler transform seems to be the winning idea:
  - very fast
  - sufficiently accurate
  - used by the newest tools (Bowtie, SOAPv2, BWA).

Review of Alignment Algorithms

- Hashing the reference genome:
  - Pros: easy to multi-thread
  - Cons: large memory footprint
- Hashing the read sequences
  - Pros: flexible memory footprint
  - Cons: difficult to multi-thread
- Alignment by merge sort:
  - Pros: flexible memory
  - Cons: hard for paired reads
- Indexing genome by Burrows-Wheeler Transform
  - Pros: fast and relatively small memory footprint
  - Cons: not applicable to long reads

Popular Alignment Tools

- Eland (Solexa)
  - supplied by Illumina as part of the Solexa Pipeline
  - very fast
  - does not make use of quality scores
- Maq (Li *et al.*, Sanger Institute)
  - widely used
  - interprets quality score and estimates alignment score
  - downstream analysis tools (SNP, indel calling)
  - can deal with SOLiD colour space data
  - being replaced by BWA
- Bowtie (Langmead *et al.*, Univ of Maryland)
  - based on Burrows-Wheeler transform
  - very fast, good accuracy
  - downstream tools available

Aligning Hashed Reads

- Naive algorithm:
  - Make a hash table of the first 28mers of each read, so that for each 28mer, we can look up quickly which reads start with it.
  - Then, go through the genome, base for base. For each 28mer, look up in the hash table whether reads start with it, and if so, add a note of the current genome position to these reads.
- Problem: What if there are read errors in the first 28 base pairs?

De Novo Assembly

- NGS offers the possibility to sequence anything and aligning the reads against "reference" genome is straightforward.
- But what if there is no such "reference" genome?
  - "de novo" assembly

De Novo Assembly

- Assembly requires specialized software, typically based on so-called de-Brujin graphs
- Most popular assembly tool:
  - Velvet (Zerbino et al.)
  - ABySS (Simpson et al.)
- Solexa reads are too short for *de novo* assembly of large genomes:
  - for prokaryotes and simple eukaryotes, reasonably large contigs can be assembled.
- Using paired-end reads with very large end separation is crucial.

Paired Read Alignment

- When aligning mate paired-end data, the aligner can use the information that mate-paired reads have a known separation:
  - Try to align the reads individually
  - Then, for each aligned read, attempt to align the mate in a small window near the first read's position with a more sensitive algorithm, e.g., Smith-Waterman to allow for gaps.
- Be sure to tell the aligner the minimal and maximal separation.
  - This allows to find small indels.

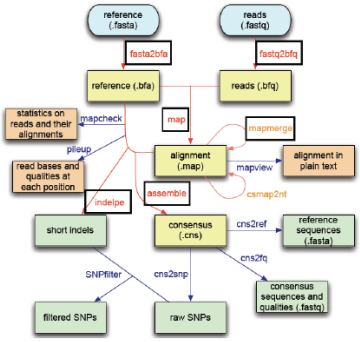
SNP Calling

- NGS is well suited for re-sequencing
  - If a base differs from the reference in most reads that are aligned to this locus, it is a likely SNP
  - If the difference occurs in half of the reads, it is a heterozygous SNP.
  - If it appears in only a few reads, it could also be a read error.
- Calculating a *p*-value for a SNP call is straightforward
  - Complication: Include base-call and alignment qualities as priors; interdependence of bases causes bias

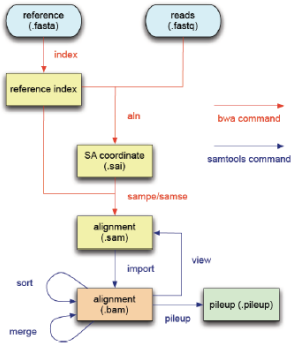
SNP Calling Software

- Some aligners com with SNP calling functionality
  - Maq
  - SOAP
  - Bowtie has a converter to Maq's format to allow to use Maq's facilities
  - For BWA, the SAMtools can be used
- Output is a list of SNPs, if possible with *p*-values
  - Due to large number of alignment software no standard modules for SNP calling have been developed.
    - SAM (the Sequence/Alignment Map format) may become a standard.

Maq Pipeline



BWA/SAMTools Pipeline



Tools for RNA-Seq

- RNA-Seq has additional challenges:
  - Reads may straddle splice junctions
  - Paralogy between genes prevent unique mappings
  - One may want to incorporate or amend known gene models
- Specialized tools for RNA-Seq alignment:
  - ERANGE
  - TopHat
  - G-Mo.R-Se
  - edgeR
  - BayesSeq

Writing Own Software

- The "glue" to combine the available tools is mostly missing.
- You will have to write your own scripts.
- Often used languages:
  - Perl
  - Python
  - R
  - Java
  - C/C++

Using R

- Pros:
  - Huge statistical library
  - Large bioinformatics library
  - Good plotting facilities
  - Convenient interactive shell
- Cons:
  - Call-by-value semantics not well suited for very large amounts of data
  - Slow due to lack of bytecode compiler
  - Poor string-handling abilities

Bioconductor NGS Packages

- Biostrings
- BSGenome
- ShortRead
- TileQC
- GenomeGraphs
- HilbertVis
- TileQC
- ChipSeq
- edgeR

Pathway to Genomic Medicine

In spite of the tremendous progress, WHY are we still very far from Personalized Medicine?

Human Genome Project	→	ENCODE Project	→	HapMap Project	→	Genomic Medicine
Sequencing of the human DNA		Interpreting the human genome sequence		Implicating genetic variants with human disease		Personalized medicine Cure for diseases