Intro 2: Last week's take home lessons

Elements & Purification

Systems Biology & Applications of Models

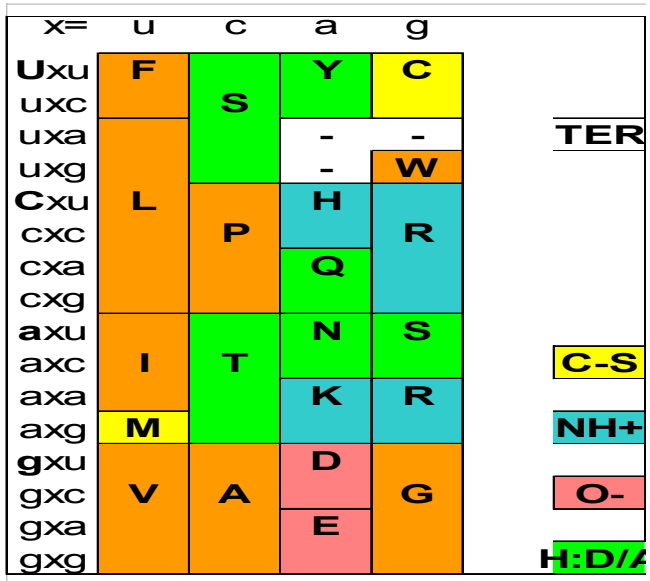
Life Components & Interconnections

Continuity of Life & Central Dogma

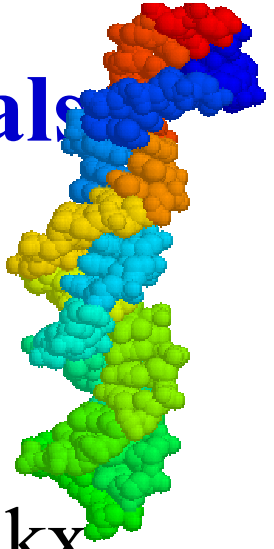
Qualitative Models & Evidence

Functional Genomics & Quantitative models

Mutations & Selection



DNA 1: Today's story, logic & goals



Types of mutants

Mutation, drift, selection

Binomial & exponential $dx/dt = kx$

Association studies χ^2 statistic

Linked and causative alleles

Haplotypes

Computing the first genome,

the second ...

New technologies

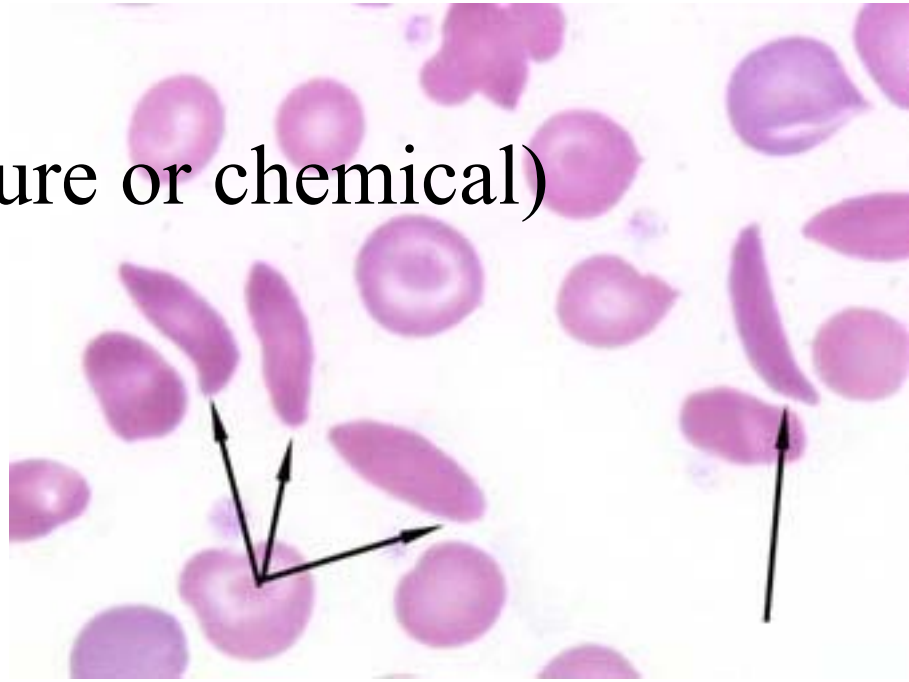
Random and systematic errors

Connecting Genotype & Phenotype

%DNA identity	
100%	Functional measures
99.9%	**S**ingle **N**ucleotide **P**olymorphisms (SNPs)
70-98%	Speciation
30%	Sequence homology
<25%	Distant (detectable only in 3D structures)

Types of phenotypic effects of mutations

Null: PKU
Dosage: Trisomy 21
Conditional (e.g. temperature or chemical)
Gain of function: HbS
Altered ligand specificity



Types of mutations

Single substitution: A to C, G or T, etc.
Deletion: 1 bp ... chromosomes (aneuploidy)
Duplication: as above (often at tandem repeats)
Inversion: ABCDEFG to AB**edc**FG
Translocation: ABCD & WXYZ to AB**YZ** & WX**CD**
Insertion: ABCD to ABινσερτCD
Recombination: ABCDEFGH & AB**c**DE**f**GH to
AB**c**DEFGH & ABCDE**f**GH

Mutations & Polymorphisms

Mutations become polymorphisms or "common alleles" when frequency > 1% in a population (arbitrary)

All Single Nucleotide Polymorphisms (SNPs) (probably) exist in the human population:
3 billion x 4 (ACGT) at frequencies near 10^{-5} .

SNPs linked to a phenotype or causative.

Haplotypes

Representation of the DNA sequence of one chromsome (or smaller segments in cis).

Indirect inference from pooled diploid data

Direct observation from meiotic or mitotic segregation, cloned or physically separated
chromsomes or segments

Linkage & Association

Family Triad: parents & child vs case-control

vs.

Case-control studies of association in structured or admixed populations. Pritchard &Donnelly, 2001. To appear in Theor. Pop. Biol. Program STRAT

Null hypothesis: allele frequencies in a candidate locus do not depend on phenotype (within subpopulations)

Pharmacogenomics

Examples of clinically relevant genetic polymorphisms influencing drug metabolism and effects.

Additional data **(http://www.sciencemag.org/feature/data/1044449.shl)**

Gene/Enzyme	Drug	Quantitative effect
CYP2C9	Tolbutamide, warfarin, phenytoin, nonsteroidal anti-inflammatories	Anticoagulant effect of warfarin
CYP2D6	Beta blockers, antidepressants, antipsychotics, codeine, debrisoquin, dextromethorphan, encainide, flecainide, guanoxan, methoxyamphetamine, *N*-propylajmaline, perhexiline, phenacetin, phenformin, propafenone, sparteine	Tardive dyskinesia from antipsychotics; narcotic side effects, efficacy, and dependence; imipramine dose requirement; beta-blocker effect
Dihydropyrimidine dehydrogenase	Fluorouracil	Fluorouracil neurotoxicity
Thiopurine methyltransferase	Mercaptopurine, thioguanine, azathioprine	Thiopurine toxicity and efficacy; risk of second cancers
ACE	Enalapril, lisinopril, captopril	Renoprotective effects, cardiac indices, blood pressure, immunoglobulin A nephropathy
Potassium channels		
HERG	Quinidine	Drug-induced long QT syndrome
	Cisapride	Drug-induced torsade de pointes
KvLQT1	Terfenadine, disopyramide, meflaquine	Drug-induced long QT syndrome
hKCNE2	Clarithromycin	Drug-induced arrhythmia

DNA Diversity Databases

~100 genomes completed (GOLD) **(http://wit.integratedgenomics.com/GOLD/)**

A list of SNP databases **(http://ariel.ucs.unimelb.edu.au/~cotton/mdi.htm)**

3 million human SNPs www.ncbi.nlm.nih.gov/SNP

mapped snp.cshl.org

23K to 60K SNPs in genes HGMD

(http://snp.cshl.org/naturepaper.html), (http://archive.uwcm.ac.uk/uwcm/mg/docs/hahaha.html)

Causative SNPs can be in non-coding repeats

aggc**A**ggtggatca

aggc**G**ggtggatca

ALU repeat found upstream of Myeloperoxidase

"severalfold less transcriptional activity"
"-463 G creates a stronger SP1 binding site & retinoic acid response element (RARE) in the allele... overrepresented in acute promyelocytic leukemia"

Piedrafita FJ, et al. 1996 *JBC 271: 14412*
(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=8662930&dopt=Abstract)

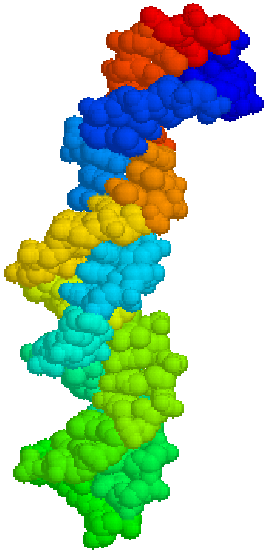
Modes of inheritance

DNA, RNA (e.g. RNAi), protein (prion),
& modifications (e.g. 5mC)

"Horizontal" (generally between species)
transduction, transformation, transgenic

"Vertical"
Mitosis: duplication & division (e.g. somatic)
Meiosis/fusion: diploid recombination, reduction
Maternal (e.g. mitochondrial)

Today's story, logic & goals



Types of mutants

Mutation, drift, selection

Binomial & exponential $dx/dt = kx$

Association studies χ^2 statistic

Linked and causative alleles

Haplotypes

Computing the first genome,

the second ...

New technologies

Random and systematic errors

Where do allele frequencies come from?

Mutation/migration(M), Selection(S), Drift (D), …

Assumptions:

Constant population size N

Random mating

Non-overlapping generations

(NOT at equilibrium, not infinite alleles,sites or N)

See: Fisher 1930, Wright 1931, Hartl & Clark 1997

(http://shop.barnesandnoble.com/textbooks/booksearch/isbnInquiry.asp?isbn=0878933069)

Directional & Stabilizing Selection

- *codominant mode of selection* (coefficient s)
  - fitness of heterozygote is the mean of the fitness(w) of the two homozygotes

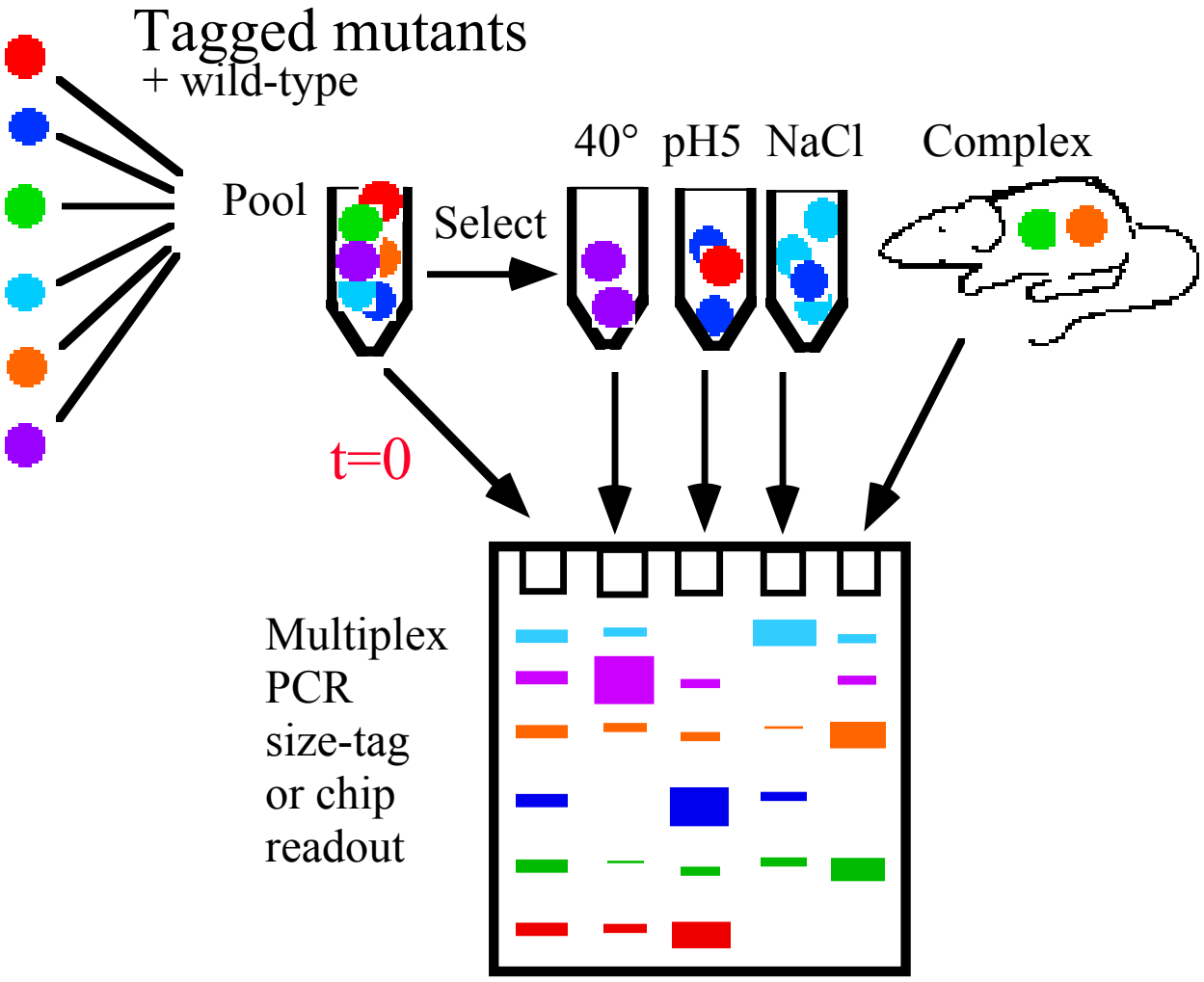
  AA = 1; Aa = 1 + s; aa = 1 + 2s
  - always increase frequency of one allele at expense of the other
- *overdominant mode*
  - heterozygote has highest fitness

  AA = 1, Aa = 1 + s; aa = 1 + t

  where $0 < t < s$
  - reach equilibrium where two alleles coexist

  See H&C 1997 p. 229

Ratio of strains over environments, **e** , times, $\mathbf{t_e}$, selection coefficients, $\mathbf{s_e}$, $R = R_o \exp[-\Sigma s_e t_e]$



Where do allele frequencies come from?

Mutation/migration(M), Selection(S), Drift (D), …

$$\mathbf{M}_j = \sum_{i=0,j}(T_i * B[N-i, j-i, \mathbf{F}]); \qquad M_j = \sum_{i=j,N}(M_i * B[i, i-j, \mathbf{R}])$$

$$\mathbf{S}_j = \sum_{i=1,j}(M_i * B[N-i, j-i, 1-1/w]) \text{ if } w>1; \quad S_j = \sum_{i=j,N-1}(M_i * B[i, i-j, 1-w]) \text{ if } w<1;$$

$$\mathbf{D}_j = \sum_{i=1,N-1} S_i * B[N, j, i/N]$$

w=relative fitness of i mutants to N-i original

T_i, M_i, D_i, S_i = frequency of i mutants in a pop. size N

F= forward mutation(or migration) probability ; R=reverse.

B(N,i,p)= Binomial = $C(N,i)\, p^i\, (1-p)^{N-i}$

(Fisher 1930, Wright 1931, Hartl & Clark 1997)

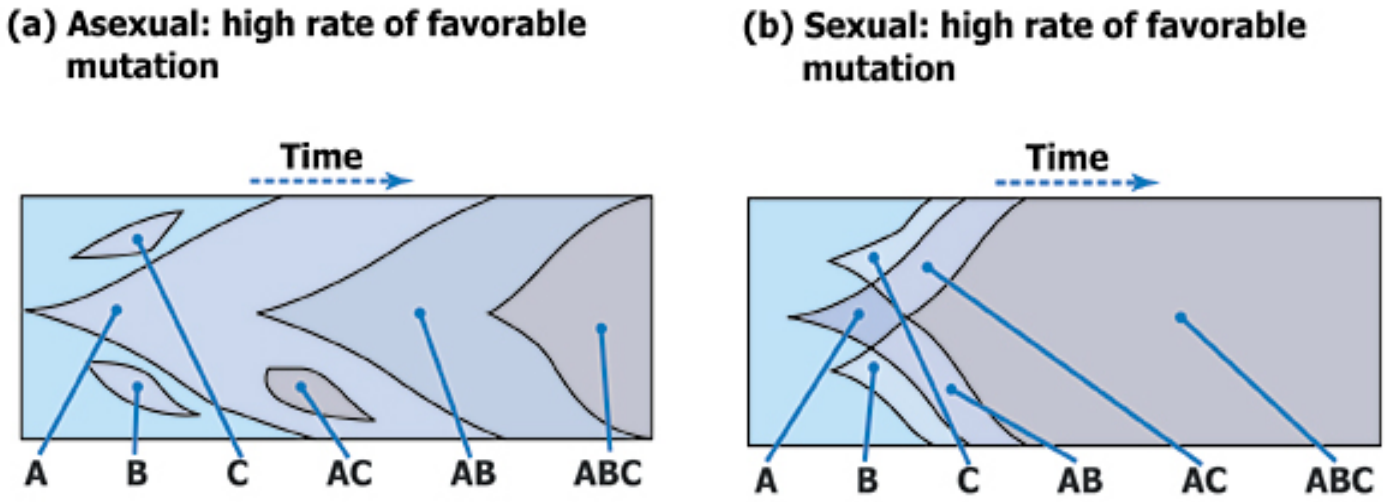
(http://shop.barnesandnoble.com/textbooks/booksearch/isbnInquiry.asp?isbn=0878933069)

Random Genetic Drift

very dependent upon population size

Role of Genetic Exchange

- Effect on distribution of fitness in the whole population
- Can accelerate rate of evolution at high cost (50%)



from Crow & Kimura 1970
Clark & Hartl 1997 p. 182

Common Disease – Common Variant Theory. How common?

ApoE allele ε4 : Alzheimer's dementia, & hypercholesterolemia
20% in humans, >97% in chimps
(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=7772071&dopt=Abstract)

HbS 17% & G6PD 40% in a Saudi sample

CCR5Δ32 : resistance to HIV
9% in caucasians

Are rare variants responsible for susceptibility to complex diseases?

"Customary in theoretical work relating to complex diseases, the allele frequencies ... are treated as parameters of the model"
New here: "resulting from an evolutionary process including selection, mutation, and genetic drift ... to learn about the underlying allele frequencies"

See Pritchard Am.J.Hum.Gen 69:124-137. (2001) Programs
(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=11404818&dopt=Abstract)

(http://www.stats.ox.ac.uk/~pritch/software.html)

Top 5 RR

DNA1: Today's story, logic & goals

Types of mutants

Mutation, drift, selection

Binomial & exponential $dx/dt = k$

Association studies χ^2 statistic

Linked and causative alleles

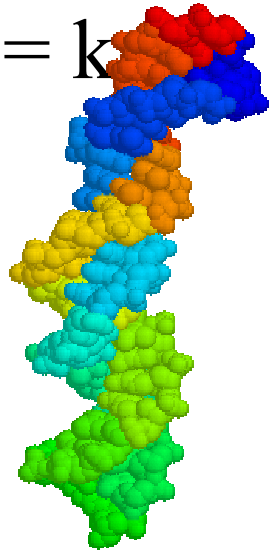
Haplotypes

Computing the first genome,

the second ...

New technologies

Random and systematic errors



One form of HIV-1 Resistance

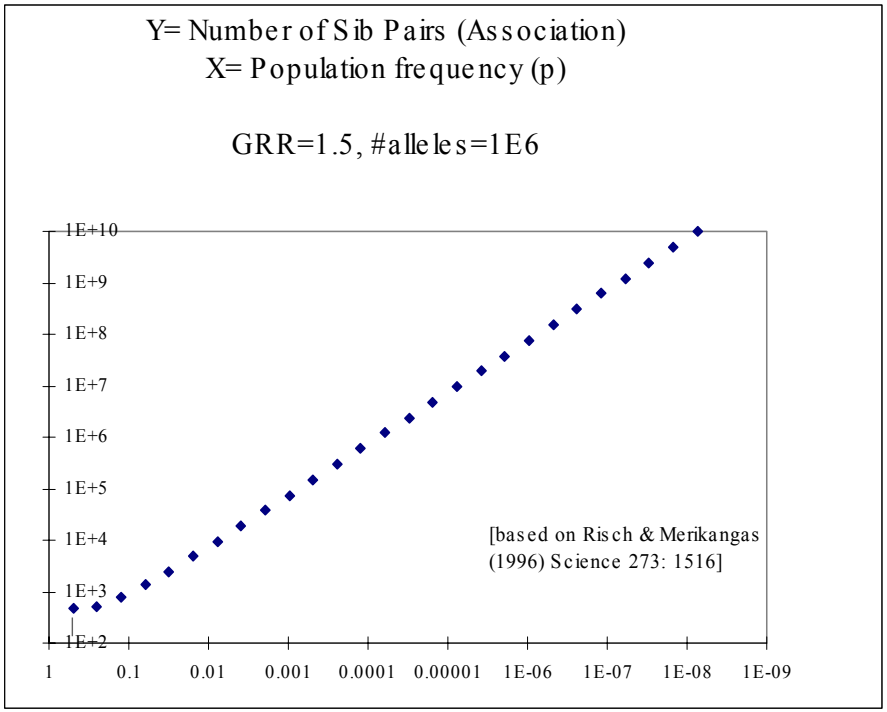
Association test for CCR-5 & HIV resistance

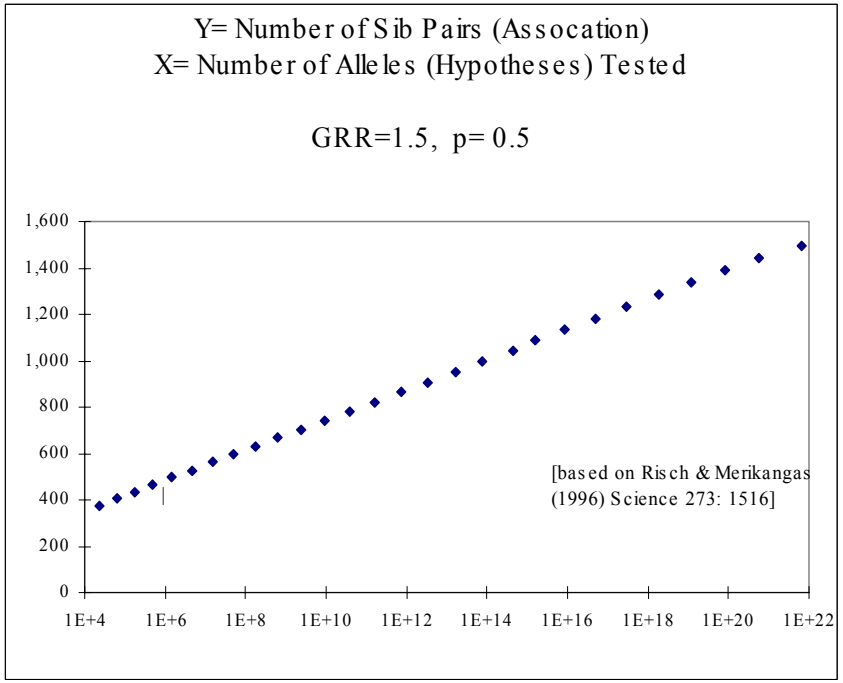
Alleles	**Obs Neg**	**ObsSeroPos**	total	**ExpecNeg**	**ExpecPos**
CCR-5+	1278	1368	2646	1305	1341
Δ ccr-5	130	78	208	103	105
total	1408	1446	2854		
					P
dof=(r-1)(c-1)=1		ChiSq=sum[(o-e)^2/e]=		**15.6**	0.00008

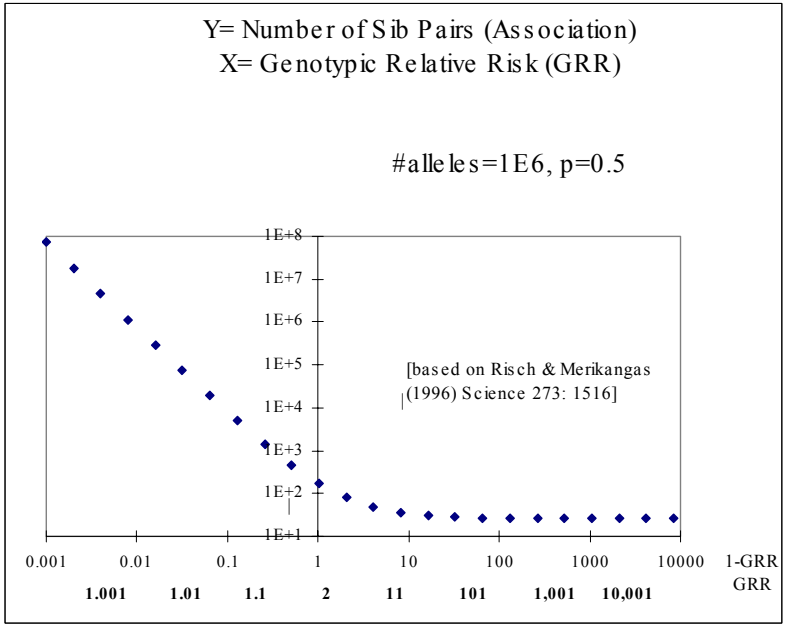
Samson et al. Nature 1996 382:722-5

(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=8751444&dopt=Abst

But what if we test more than one locus?







The future of genetic studies of complex human diseases.

Ref

(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=8801636&dopt=Abstract)

GRR = Genotypic relative risk

How many "new" polymorphisms?

G= generations of exponential population growth = 5000

N'= population size = 6×10^9 now; N= 10^4 pre-G

m= mutation rate per bp per generation = 10^{-8} to 10^{-9} (ref)

(http://www.nature.com/cgi-taf/DynaPage.taf?file=nature/journal/v397/n6717/abs/397344a0_fs.html&filetype=&content_filetype=)

L= diploid genome = 6×10^9 bp

$e^{kG} = N'/N$; so $k = 0.0028$

Av # new mutations per genome $< \sum_{t=1 \text{ to } 5000} Le^{kt}m = 4 \times 10^3$ to 4×10^4

Take home: "High genomic deleterious mutation rates in hominids" accumulate over 5000 generations & confound linkage methods And common (causative) allele assumptions.

Finding & Creating mutants

Isogenic
Proof of causality:
Find > Create a copy > Revert

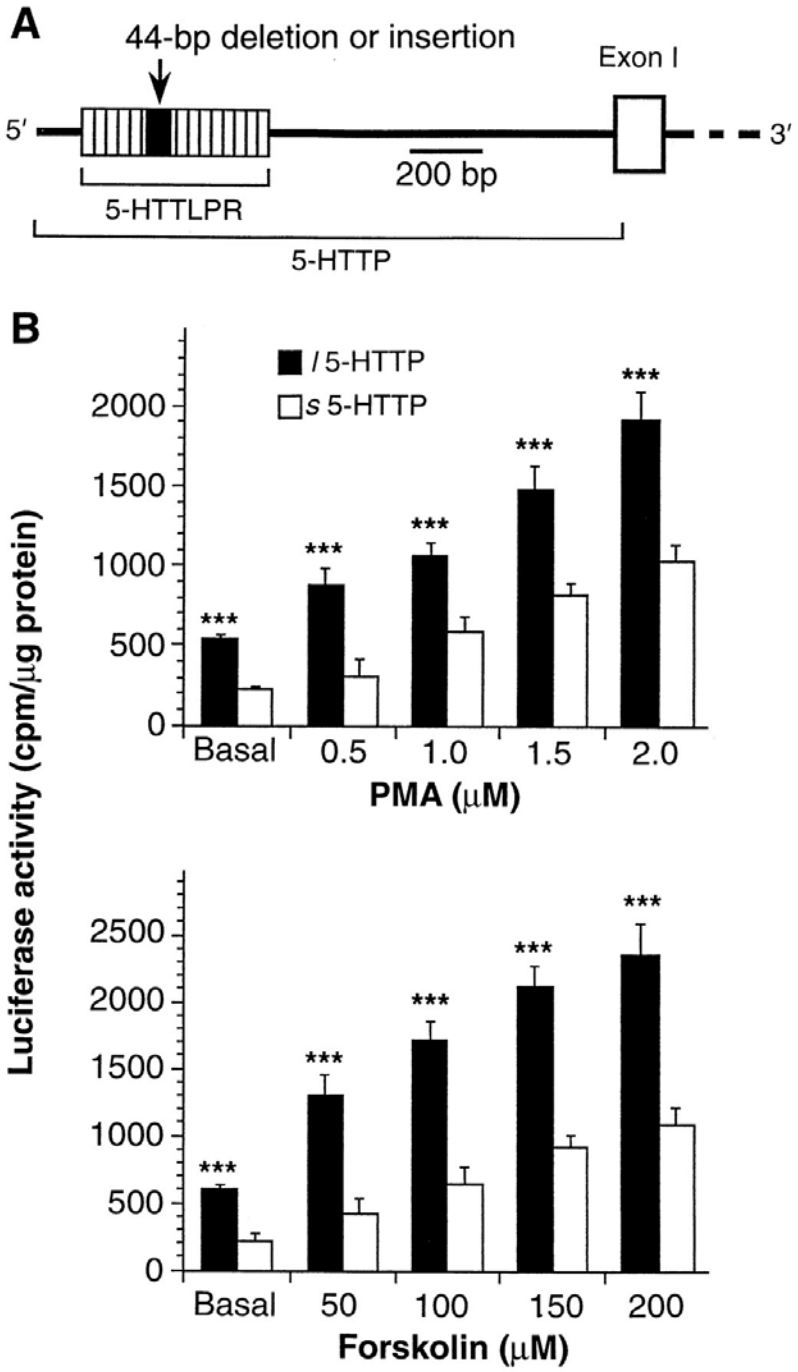
Caution:
Effects on nearby genes
Aneuploidy (ref)

(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=10888885&dopt=Abstract)

Pharmacogenomics

Example

5-hydroxytryptamine transporter



Lesch KP, et al Science 1996 274:1527-31
Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. Pubmed
(http://www.ncbi.nlm.nih.gov/htbinpost/Entrez/query?uid=8929413&form=6&db=m&Dopt=b)

Caution: phases of human genetics

Monogenic vs. Polygenic dichotomy

Method	Problems
Mendelian Linkage (300bp)	need large families
Common indirect/LD (10^6bp)	recombination & new alleles
Common direct (causative)	3% coding + ?non-coding
All alleles (10^9)	expensive ($0.20 per SNP) (methods)

(http://www.ncbi.nlm.nih.gov/SNP/snp_tableList.cgi?type=method)

DNA1: Today's story, logic & goals

Types of mutants

Mutation, drift, selection

Binomial & exponential $dx/dt = kx$

Association studies χ^2 statistic

Linked and causative alleles

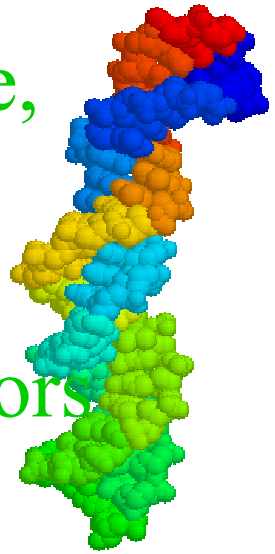
Haplotypes

Computing the first genome,

the second ...

New technologies

Random and systematic errors



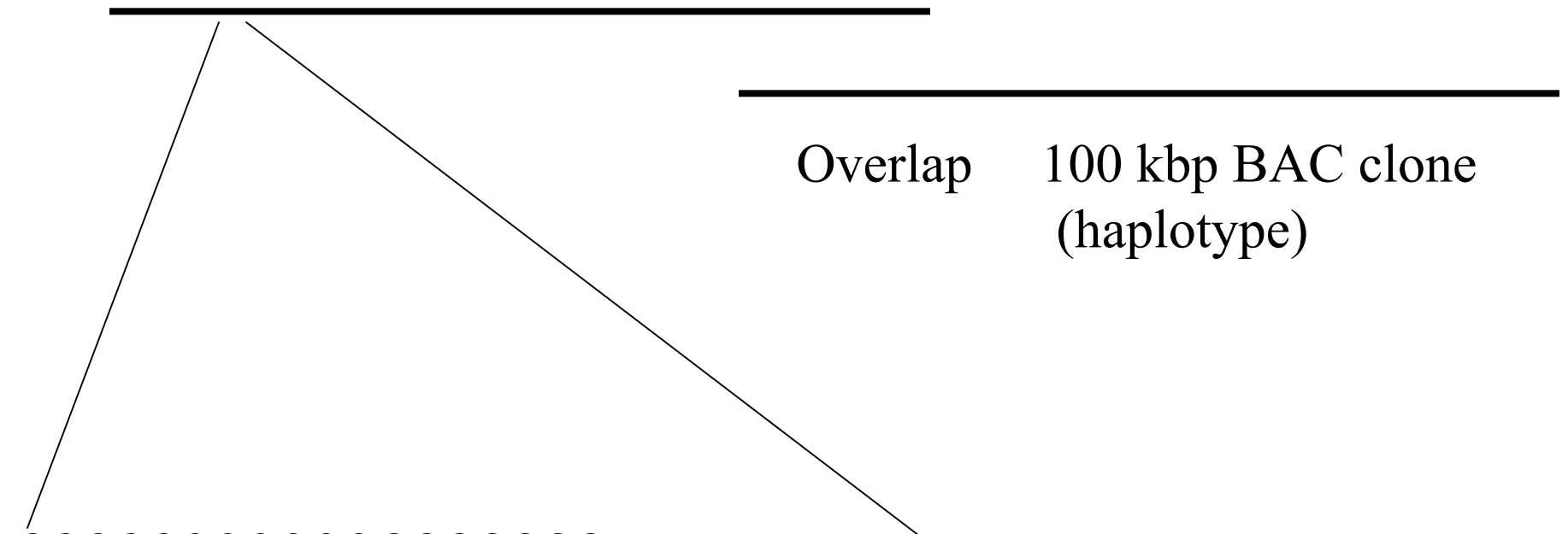
Examples of random & systematic errors?

For (clone) template isolation:

For sequencing:

For assembly:

Sequence assembly



4 sequences in 2 islands

acgacatagcgactagcta

See Ewing, Hillier, Wendl, & Green 1998

Examples of random & systematic errors?

For (clone) template isolation:

For sequencing:

For assembly:

Examples of systematic errors

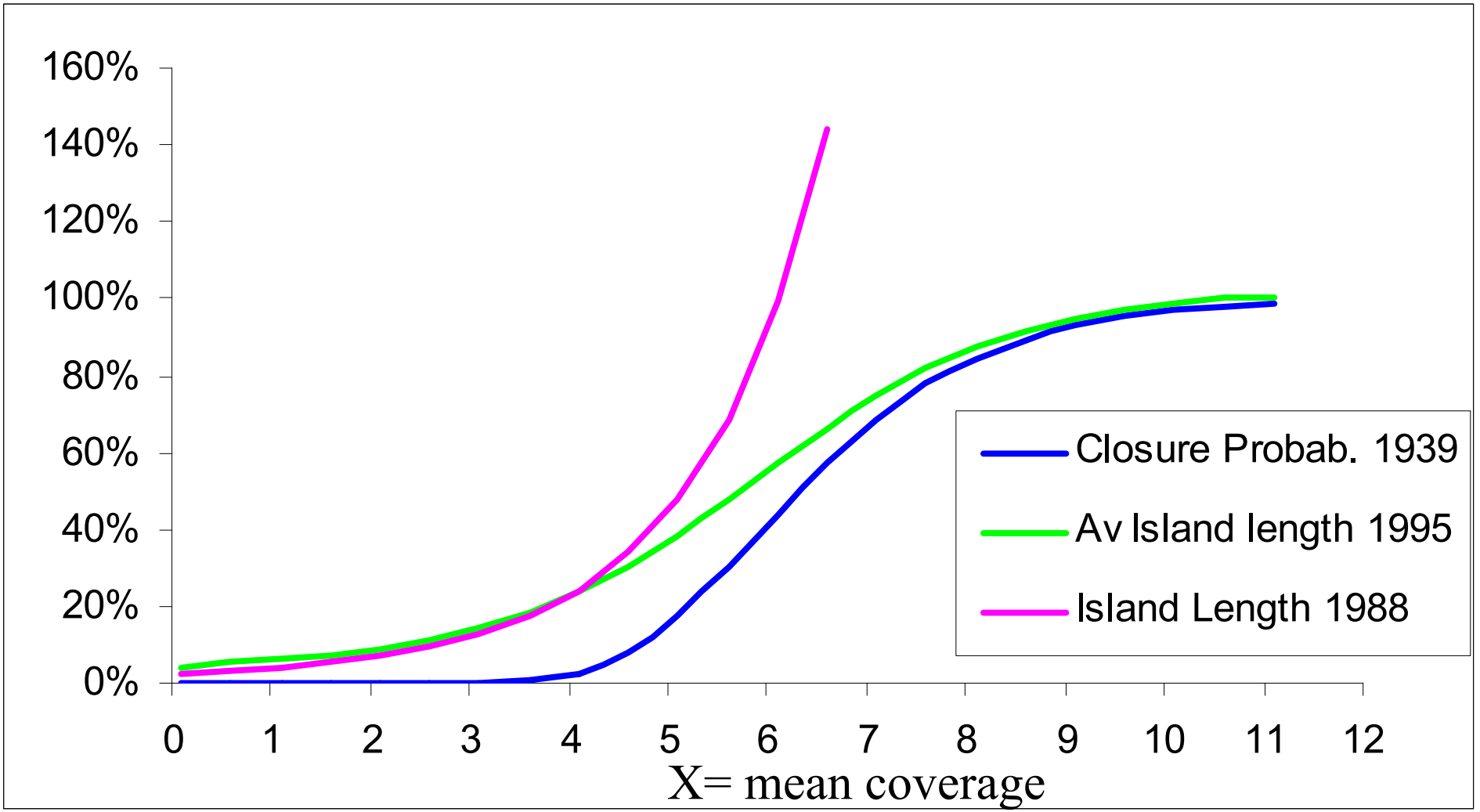
For (clone) template isolation:
restriction sites, repeats

For sequencing:
Hairpins, tandem repeats

For assembly:
repeats, errors, polymorphisms,
chimeric clones, read mistracking

Whole-genome shotgun

Project completion % vs coverage redundancy



(Roach 1995)(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=8808467&dopt=Abstract)

Table 2. Simulation Default Parameters

35-nucleotide overlap required for sequence joining
10-fold genome coverage
400-nucleotide read lengths
15% variation in insert sizes
10,000-nucleotide average size for long inserts
700-nucleotide average size for short inserts
1:1 ratio of long to short inserts
100 kb spacing between STSs
300-nucleotide STS length
20% of genome comprised of SINEs with 300-nucleotide lengths
5% of genome comprised of LINEs with 1500-nucleotide lengths
4:1 ratio of SINEs to LINEs

Weber & Myers 1997

2-Oct-2002 Boston GSAC Panel Discussion

"The Future of Sequencing Technology: Advancing Toward the $1,000 Genome"

Moderators:

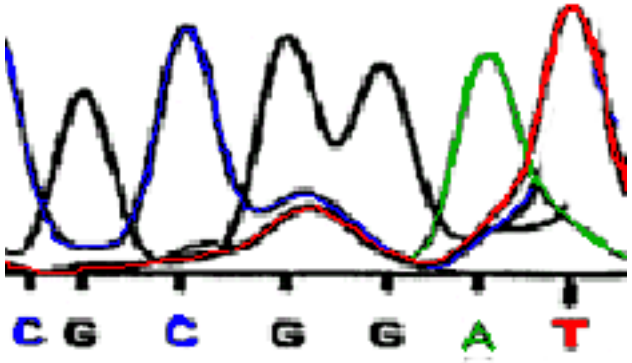
•J. Craig Venter, Ph.D., The Center for Advancement of Genomics
•Gerald Rubin, Ph.D., Howard Hughes Medical Institute

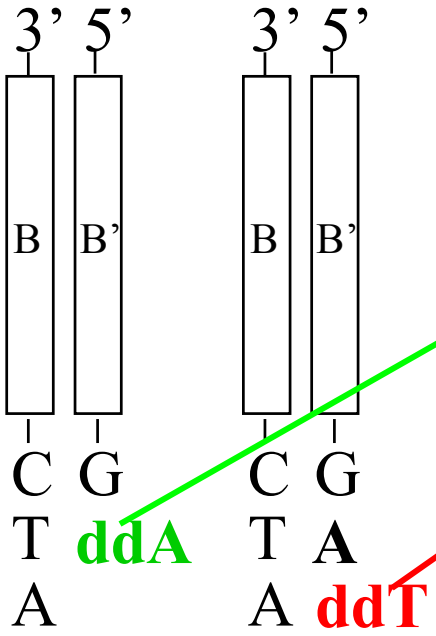
Speakers:

•George Church, Ph.D., Harvard University
•Eugene Chen, Ph.D., US Genomics
•Tony Smith, Ph.D., Solexa
•Trevor Hawkins, Ph.D., Amersham Biosciences Corporation
•Susan Hardin, Ph.D., VisiGen Biotechnologies, Inc.
•Michael P. Weiner, 454 Corporation
•Daniel H. Densham, Mobious Genomics, Ltd

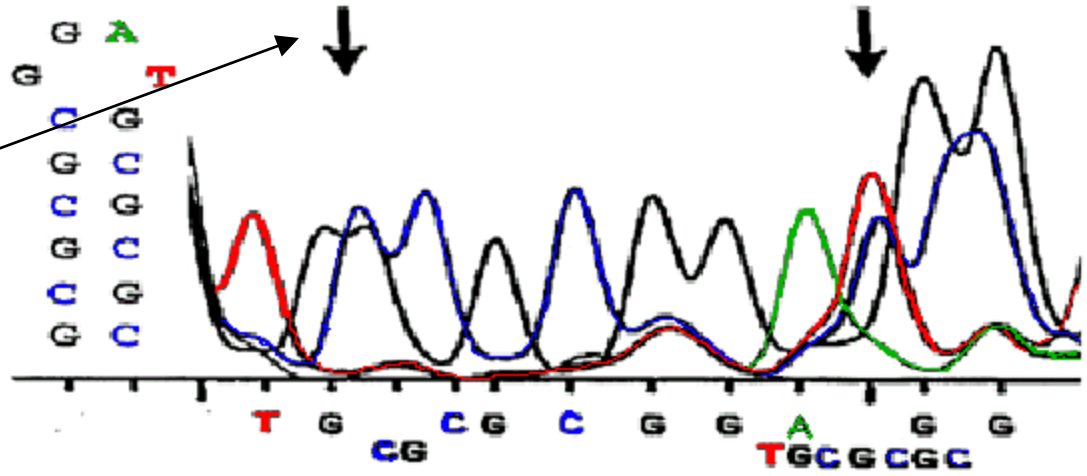
Conventional dideoxy gel with 2 hairpin

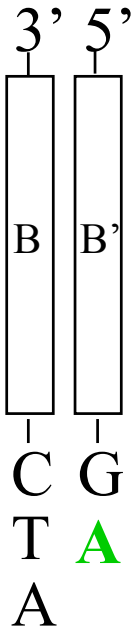
Gel size separation

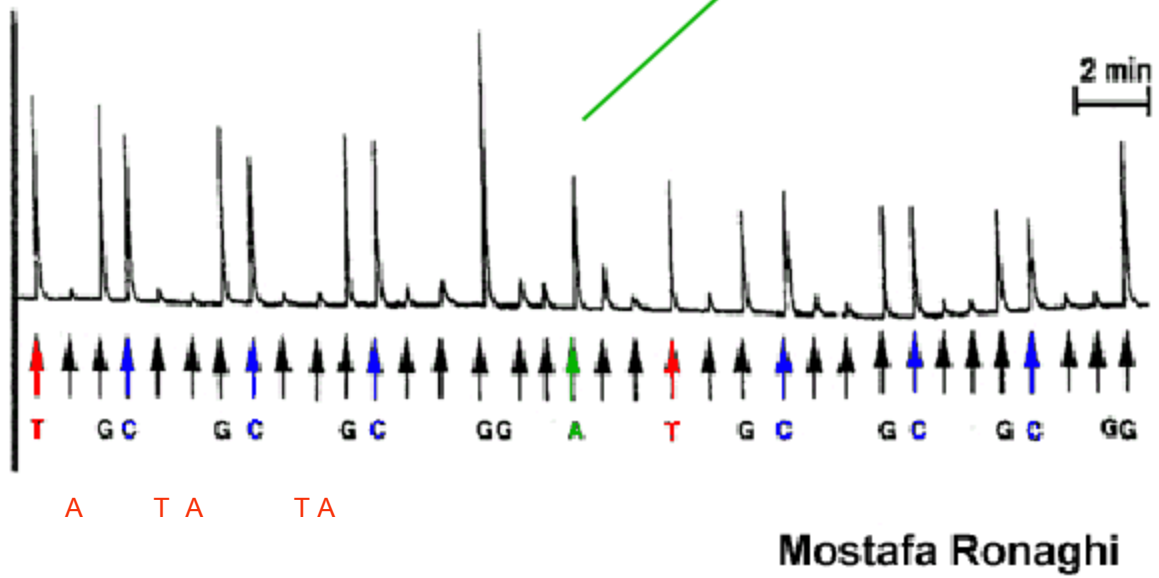




Conventional dideoxy gel with 2 hairpin Systematic errors



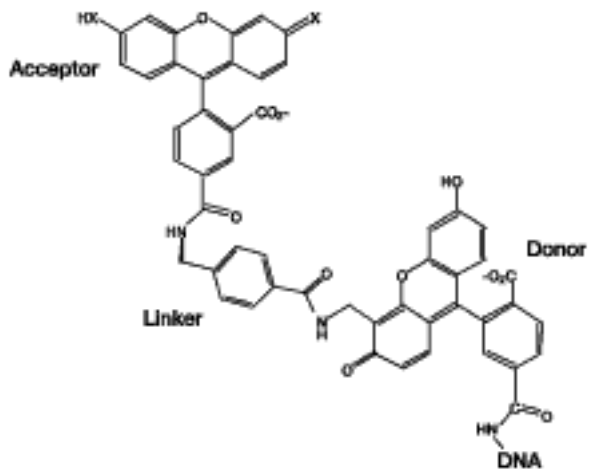


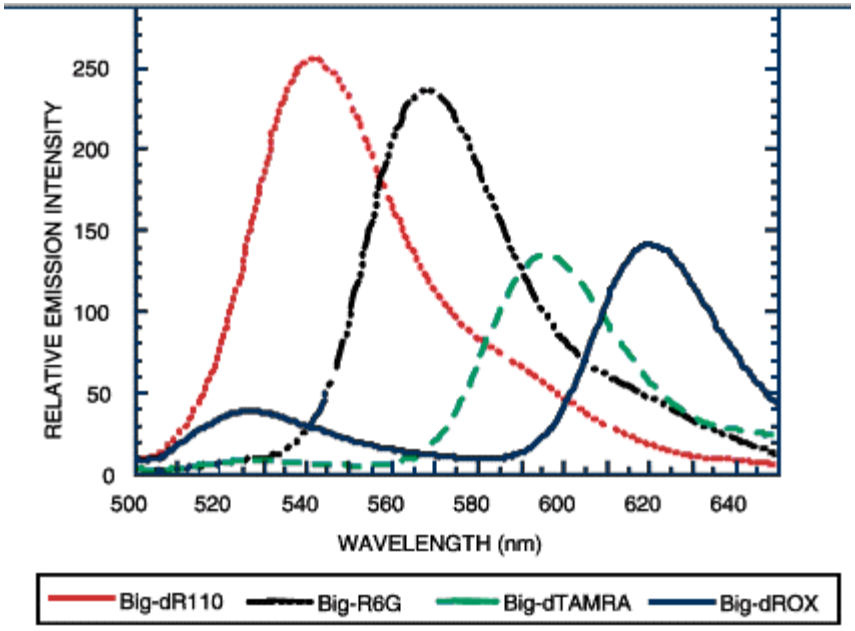


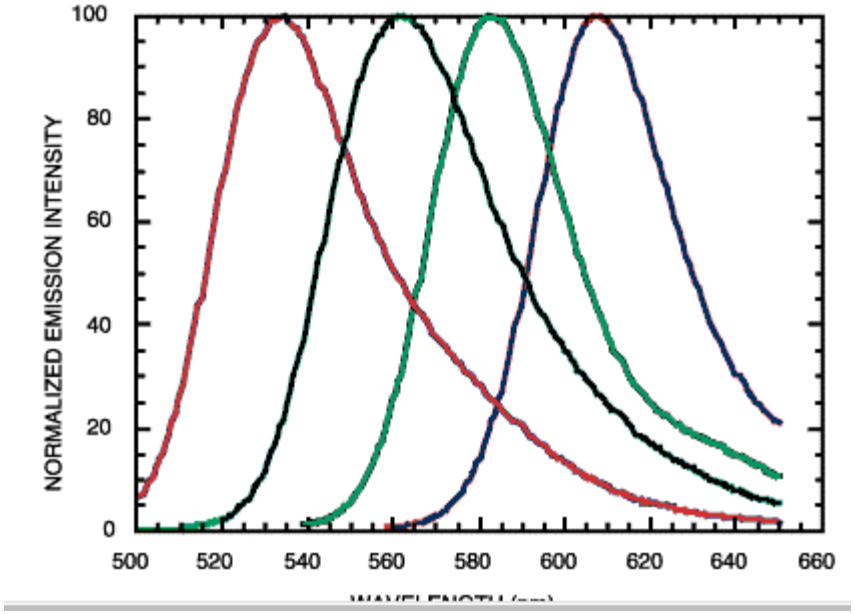
Mostafa Ronaghi

Sequential dNTP addition (pyrosequencing)
> 30 base reads; no hairpin artefacts

Fluorescent primers or ddNTPs







Anal Biochem 1997 Oct 1;252(1):78-88
Optimization of spectroscopic and electrophoretic properties of energy transfer primers.
Hung SC, Mathies RA, Glazer AN

http://www.pebio.com/ab/apply/dr/dra3b1b.html

New Genotyping & haplotyping technologies

***de novo* sequencing > scanning > selected sequencing > diagnostic methods**

Sequencing by synthesis

- **1-base Fluorescent, isotopic or Mass-spec* primer extension** (Pastinen97)
- **30-base extension Pyrosequencing** (Ronaghi99)*
- **700-base extension, capillary arrays dideoxy*** (Tabor95, Nickerson97, Heiner98)

SNP & mapping methods

- **Sequencing by hybridization on arrays** (Hacia98, Gentalen99)*
- **Chemical & enzymatic cleavage:** (Cotton98)
- **SSCP, D-HPLC (**Gross 99**)**

Femtoliter scale reactions (10^5 molecules)

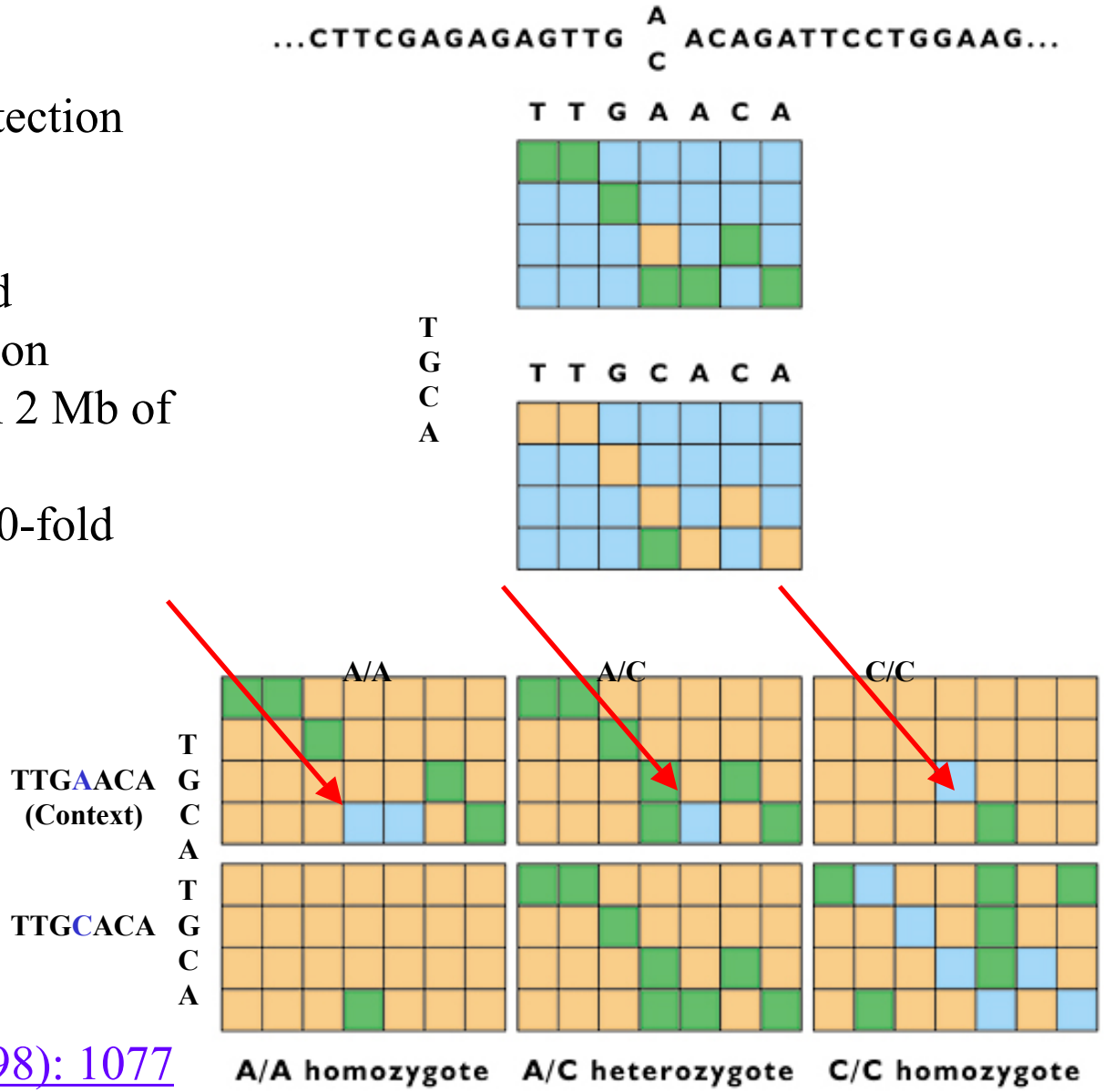
- **20-base restriction/ligation MPSS (**Gross 99**)**
- **30-base fluorescent in situ amplification sequencing (Mitra 1999)**

Single molecule methods (not production)

- **Fluorescent exonuclease** (Davis91)
- **Patch clamp current during ss-DNA nanopore transit (**Kasianowicz96**)**
- **Electron, STM, optical microscopy (**Lagutina96, Lin99**)**

Use of DNA Chips for SNP ID & Scoring

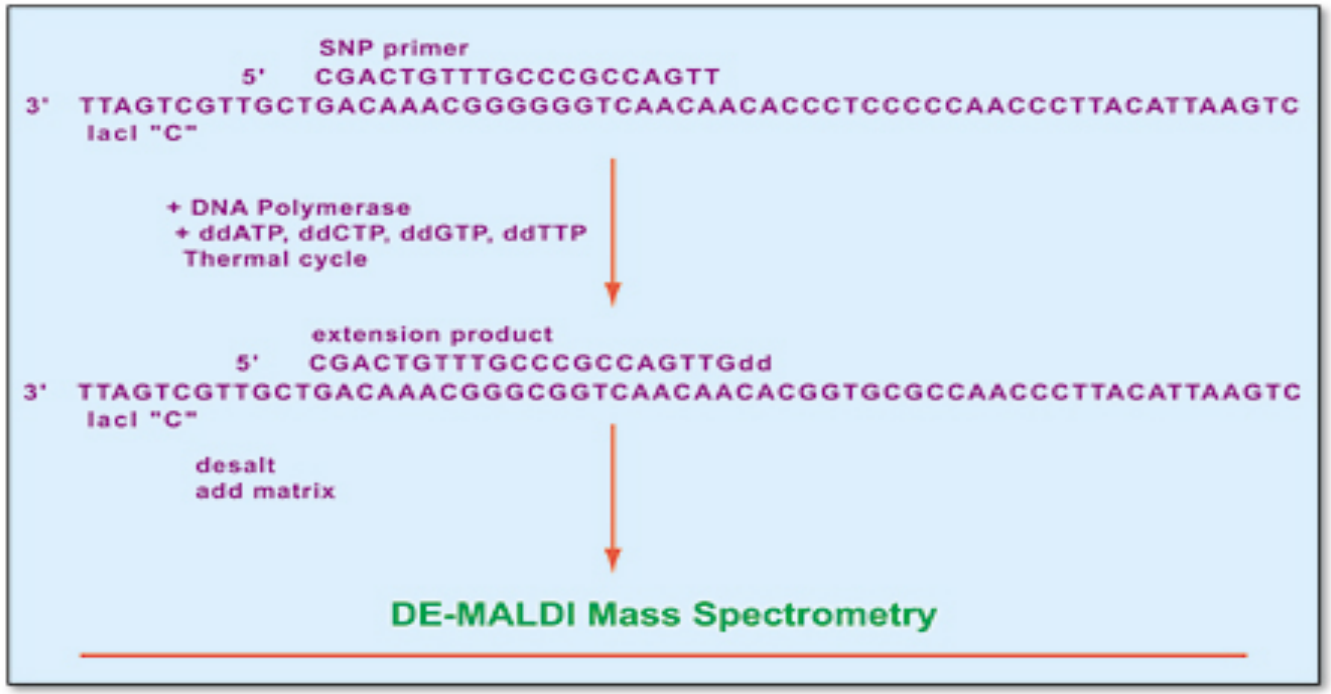
- Used for mutation detection with HIV-1, BRCA1, mitochondria
- higher throughput and potential for automation
- ID of > 2000 SNPs in 2 Mb of human DNA
- Multiplex reactions 50-fold



Wang et al., Science 280 (1998): 1077
(http://www.sciencemag.org/cgi/content/full/280/5366/1077)

Use of Mass Spec for Analysis and Scoring

Haff and Smirnov, Genome Research 7 (1997): 378

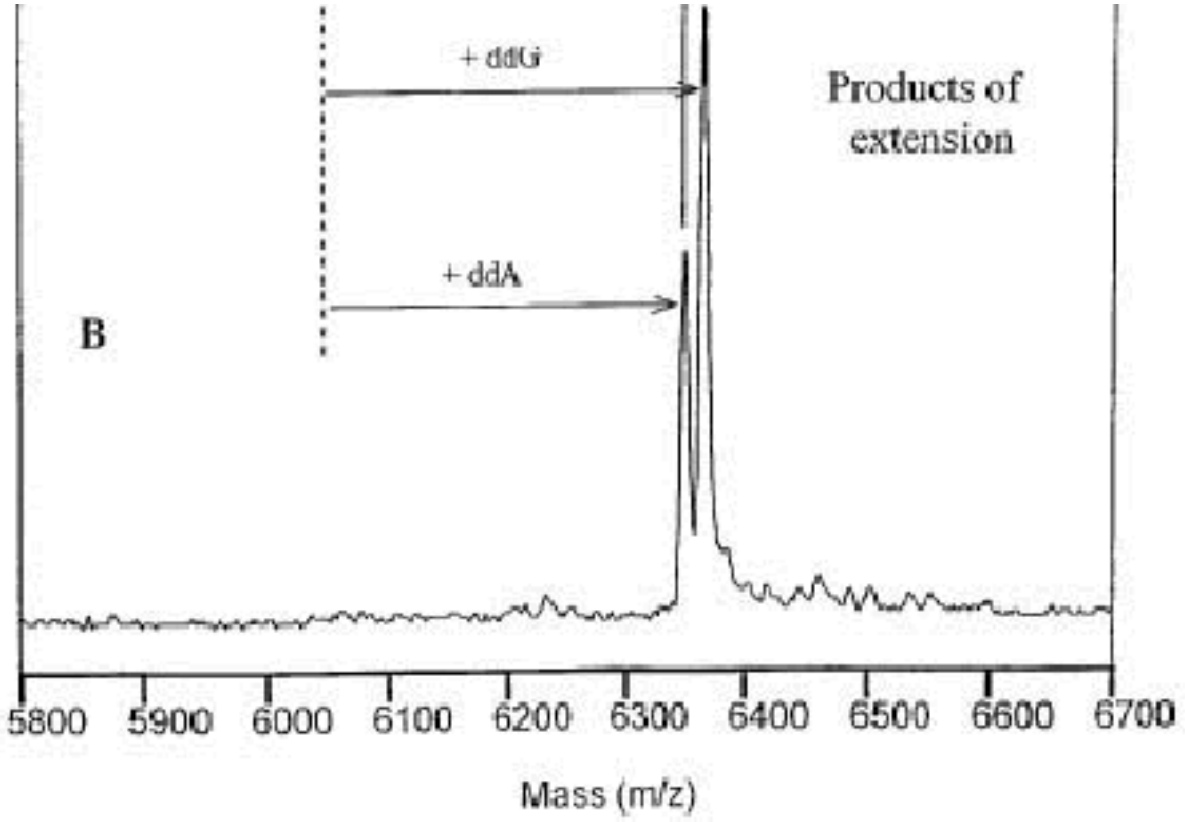


A single nucleotide primer extension assay

Mass Spectrometry for Analysis and Scoring

Haff and Smirnov, Genome Res. 7 (1997): 378

Use mass spec to score which base(s) add
Multiplex 5 with known primer masses
Pool 50 to 500 samples



Sequenom

(http://www.sequenom.com/genotyping/overview/techno/techno.html)

Searching for (nearly) exact matches

Hash
Suffix arrays
Suffix trees

$4^N \sim$ = Genome length
N=word length (for "lookup")
e.g. Set aside space for
$4^{16} \sim = 4$ billion genomic positions (each requires 4-bytes of storage).

Exact Sequence Searching

```
#!/usr/local/bin/perl
$dnatext = "gggggggCgggCgggCgggg";
print " Original genome: $dnatext \n";
$n_mut = $dnatext =~ s/gC/gg/gi;
print " Found: $n_mut mutation(s)\n";
print " After gene-therapy: $dnatext \n";
```

```
Original genome: gggggggCgggCgggCgggg
 Found: 3 mutation(s)
 After gene-therapy: gggggggggggggggggggg
```

DNA 1: Today's story, logic & goals

Types of mutants
Mutation, drift, selection
    Binomial & exponential $dx/dt = kx$
Association studies χ^2 statistic
Linked and causative alleles
Haplotypes
Computing the first genome,
    the second ...
New technologies
Random and systematic errors