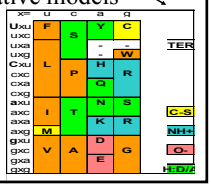
Integrate 2: Last week's take home lessons

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

2

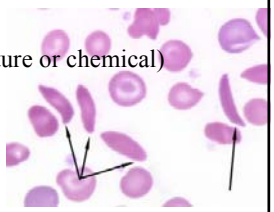
Connecting Genotype & Phenotype

%DNA identity
100% Functional measures
99.9% **S**ingle **N**ucleotide **P**olymorphisms (SNPs)
70-99% Speciation
30% Sequence homology
<25% Distant (detectable only in 3D structures)

3

Types of phenotypic effects of mutations

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



4

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 AB**ινσερτ**CD
Recombination: ABCDEFGH & AB**c**DE**f**GH to
AB**c**DEFGH & ABCDE**f**GH

5

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.

6

Mutation rates

Achondroplasia (autosomal dominant trait) FGFR3 G1138A mutations occur at 1.4 x 10^{-5} per generation)
http://www.faseb.org/genetics/ashg00/f2293.htm

Spontaneous mutation rate = 0.5 to 12 x 10^{-9} (also Anagnostopoulos *et al*. 1999; Nachman & Crowell 2000).
Frequency of induced mutations =3.4 to 90 x 10^{-9} per bp.
Weinberg, et al. 2001 Proc R Soc Lond B Biol Sci. 268(1471):1001-5.
Very high mutation rate in offspring of Chernobyl accident liquidators.

Vertebrate brain size evolution

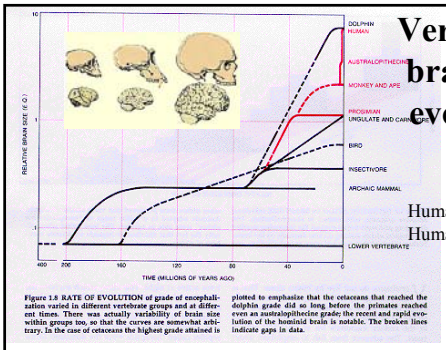


Figure 1.8 RATE OF EVOLUTION of grade of encephalization varied in different vertebrate groups and at different times. There was actually variability of brain size within groups too, so that the curves are somewhat arbitrary. In the case of cetaceans the highest grade attained is plotted to emphasize that the cetaceans that reached the dolphin grade did so long before the primates reached even an australopithecine grade; the recent and rapid evolution of the hominid brain is notable. The broken lines indicate gaps in data.

Human-chimp 1.2%
Human-human 0.1%

http://www.genome.wustl.edu/projects/chimp/ Science. 2002 Jul 19;297(5580):365-9.
Bond et al 2002 ASPM is a major determinant of cerebral cortical size. Nat Genet. 32(2):316-20.
Jerison, Paleoneurology & the Evolution of Mind, Scientific Amer. 1976

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

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
KvLQT1	Cisapride Terfenadine, disopyramide, mefloquine	Drug-induced torsade de pointes Drug-induced long QT syndrome
hKCNE2	Clarithromycin	Drug-induced arrhythmia

DNA Diversity Databases

~100 genomes completed (GOLD)

A list of SNP databases

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

mapped snp.cshl.org

23K to 60K SNPs in genes HGMD

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

13

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)

14

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

15

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

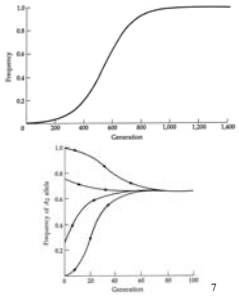
16

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

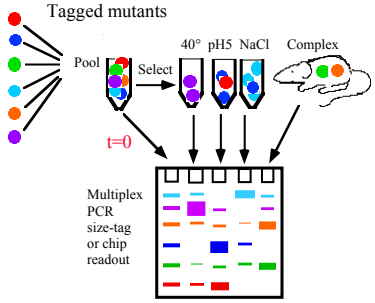
  AA = 1, Aa = 1 + s; aa = 1 + t
  where 0 < t < s
  - reach equilibrium where two alleles coexist



H&C 1997 p. 229

7

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



18

Where do allele frequencies come from?

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

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

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

$\mathbf{D}_j = \Sigma_{i=1,N-1} S_i * B[N,j,i/N]$ $\mathbf{T}_j = D_j$ (& iterate)

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.

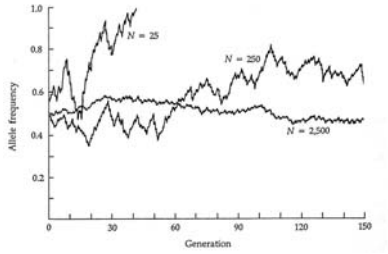
$B(N,i,p)$= Binomial = $C(N,i)\ p^i\ (1-p)^{N-i}$

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

19

Random Genetic Drift

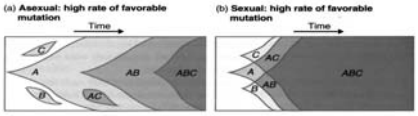
very dependent upon population size



20

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.2182

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

22

Common Disease – Common Variant Theory. How common?

ApoE allele ε4 : Alzheimer's dementia, & hypercholesterolemia
20% in humans, >97% in chimps

HbS 17% & G6PD 40% in a Saudi sample

CCR5Δ32 : resistance to HIV
9% in caucasians

23

One form of HIV-1 Resistance

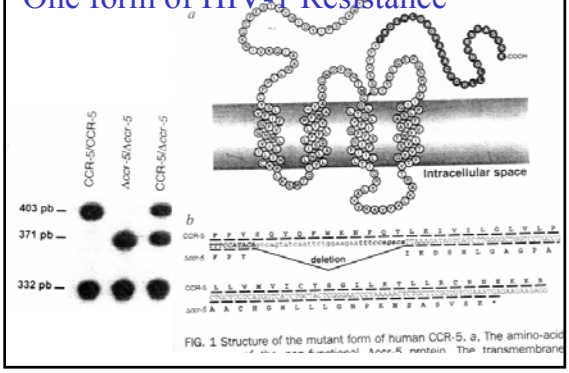
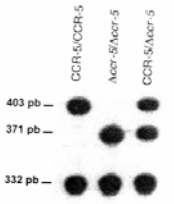


FIG. 1 Structure of the mutant form of human CCR-5. a, The amino-acid

Association test for CCR-5 & HIV resistance

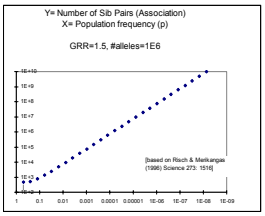
Alleles	Obs Neg	ObsSeroPos	total	ExpecNeg	ExpecPos
CCR-5+	1278	1368	2646	1305	1341
Δ ccr-5	130	78	208	103	105
total	1408	1446	2854		
					P
dof=(r-1)(c-1)=1		ChiSq=sum[(o-e)^2/e]=		**15.6**	0.00008

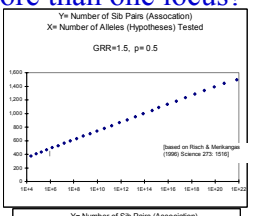


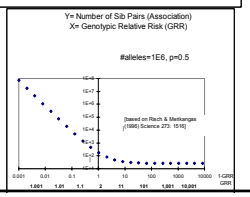
Samson et al. Nature 1996 382:722-5

25

But what if we test more than one locus?







The future of genetic studies of complex human diseases. ref

GRR = Genotypic relative risk

How many "new" mutations?

G= generations of exponential population growth = 5000
N'= population size = 6 x 10^9 now; N= 10^4 pre-G
m= mutation rate per bp per generation = 10^{-8} to 10^{-9} (ref)
L= diploid genome = 6 x 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.

27

Finding & Creating mutants

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

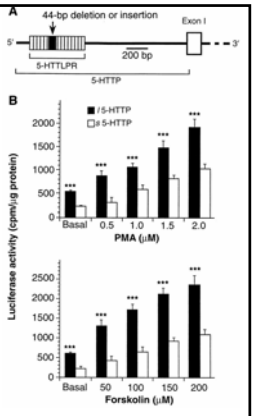
Caution:
Effects on nearby genes
Aneuploidy (ref)

28

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

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)

30

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 error

31

Why improve beyond current 1kbp/$?

Human genomes (6 billion)^2 = 10^{19} bp
Immune & cancer genome changes >10^{10} bp per time point
RNA ends & splicing: ***in situ*** 10^{12} bits/mm^3
Biodiversity: Environmental & lab evolution
Compact storage 10^5 now to 10^{17} bits/ mm^3 eventually

& How? **($1K per genome, 10^8-10^{13} bits/$)**

The issue is not speed, but integration.
Cost per 99.99% bp : Including Reagents, Personnel, Equipment/5yr, Overhead/sq.m

- Sub-mm scale : 1μm = femtoliter (10^{-15})
- Instruments should match GHz / $2K CPU

32

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)

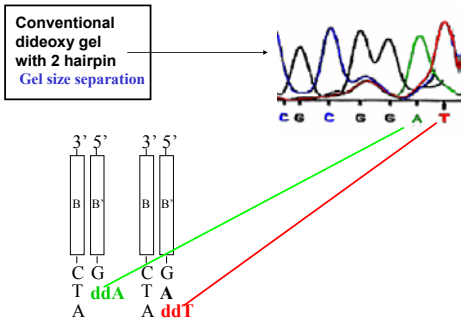
Femtoliter scale reactions (10^5 molecules)

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

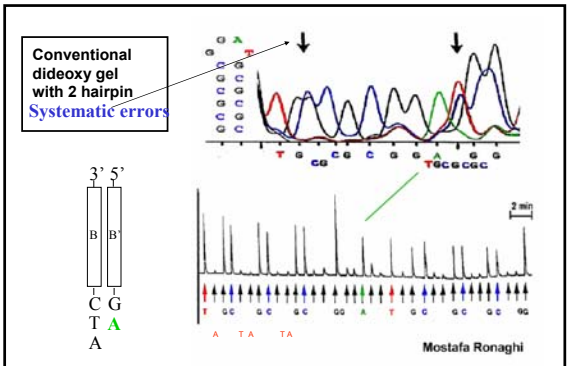
Single molecule methods (not production)

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

33



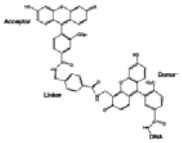
34

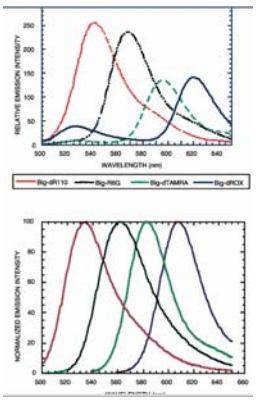


Sequential dNTP addition (Pyrosequencing) > 30 base reads; no hairpin artefacts

35

Fluorescent primers or ddNTPs

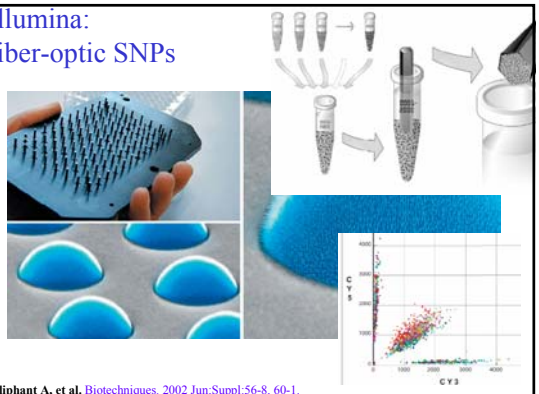




Anal Biochem 1997 Oct 1;252(1):78-88
Optimization of spectroscopic and electrophoretic properties of energy transfer primers.
Hung SC, Mathies RA, Glazer AN

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

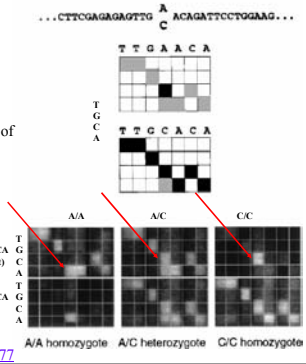
Illumina: fiber-optic SNPs



Oliphant A, et al. Biotechniques. 2002 Jun;Suppl:56-8, 60-1.
BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping.

Use of DNA Chips for SNP ID & Scoring

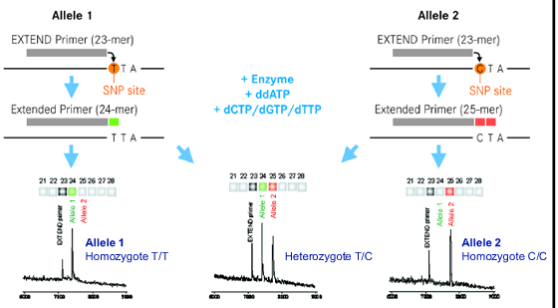
- Used for mutation detection with HIV-1, BRCA1, mitochondria
- higher throughput and potential for automation
- ID of > 2000 SNPs in 2 Mb of human DNA
- Multiplex reactions 50-fold



Kennedy et al. 2003 Nat Biotechnol. Large-scale genotyping of complex DNA.

Wang et al., Science 280 (1998): 1077

Mass Spectrometry for DNA SNPs



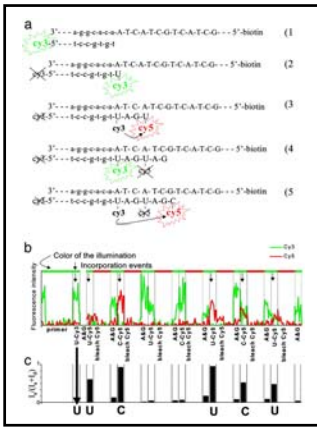
Sequenom Multiplex 5 primers Pool 50 to 500 samples Haff & Smirnov, Genome Res. 7 (1997): 378

Why single molecules?

(1) Integrate from cells/genomes/RNAs to data

(2) Geometry, "cis-ness" on a molecule, complex, or cell.
e.g. DNA Haplotypes & RNA splice-forms

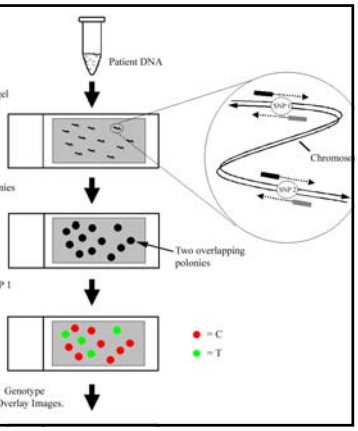
(3) Asynchronous dNTP incorporation



"Sequence information can be obtained from single DNA molecules."

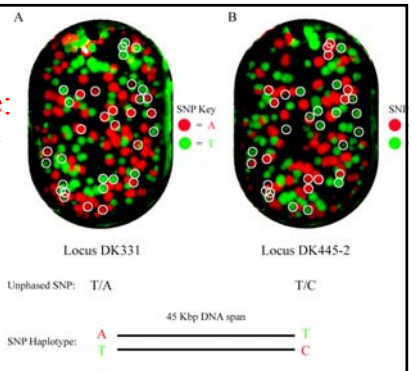
Braslavsky et al. 2003 PNAS. 100(7):3960-4.

Polymerase colonies (Polonies) along a DNA or RNA molecule



HMS: Shendure, Zhu, Butty, Williams
Wash U: Mitra
Amberger: Olejnik
U. Del: Edwards, Merritt

Human Haplotype: CFTR gene 45 kbp



Rob Mitra
Vincent Butty
Jay Shendure
Ben Williams

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).

44

Examples of random & systematic errors?

For (clone) template isolation:

For sequencing:

For assembly:

45

Examples of systematic errors

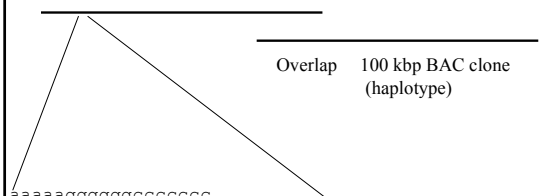
For (clone) template isolation:
restriction sites, repeats

For sequencing:
Hairpins, tandem repeats

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

46

Sequence assembly



4 sequences in 2 islands

47

acgacatagcgactagcta

Ewing, Hillier, Wendl, & Green 1998

Indel=I+D
Total= I+D+N+S

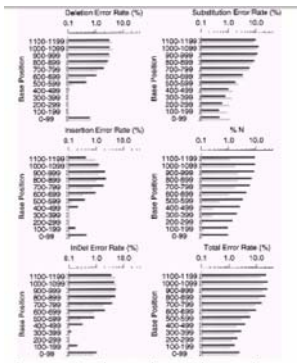
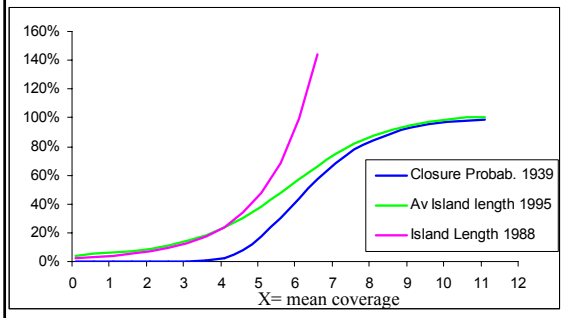


Figure 4 *Pmed* (gray bars) and ABI (black bars) error rates for dye terminator data in cosmid set 3.

Whole-genome shotgun

Project completion % vs coverage redundancy



(Roach 1995)

49

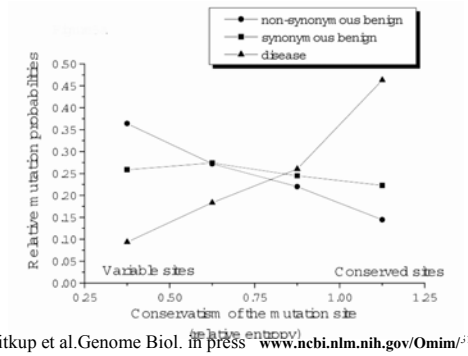
Table 2. Simulation Default Parameters

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

50

Mutable & deleterious positions



Vitkup et al.Genome Biol. in press **www.ncbi.nlm.nih.gov/Omim/**[31]

Detecting positive selection

If molecular evolution is neutral, then the ratio of amino-acid (A) to synonymous (S) polymorphism should, on average, equal that of divergence. A comparison of the A/S ratio of polymorphism in D. melanogaster with that of divergence from D. simulans shows that the A/S ratio of divergence is twice as high– since it is limited to only a fraction of the genes, which are also evolving more rapidly, this implies that positive selection is responsible.

McDonald & Kreitman Nature. 1991 Jun 20;351(6328):652-4.
Fay, Wyckoff & Wu 2002, Nature 415: 1024-1026
Smith & Eyre-Walker 2002, Nature 415:1022-4.

52

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