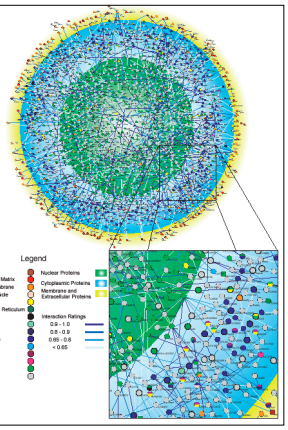
A Disease-Centric Draft Map of the Human Interactome

Joel S. Bader

Dept. of Biomedical Engineering &
High-Throughput Biology Center
Johns Hopkins University & School of Medicine
joel.bader@jhu.edu
www.jhubiomed.org

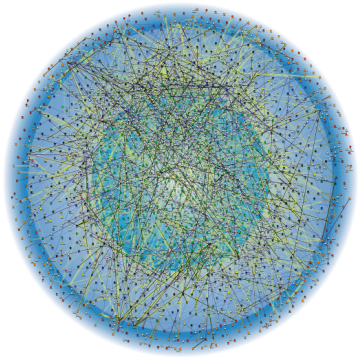
AIChE
Nov. 1, 2005



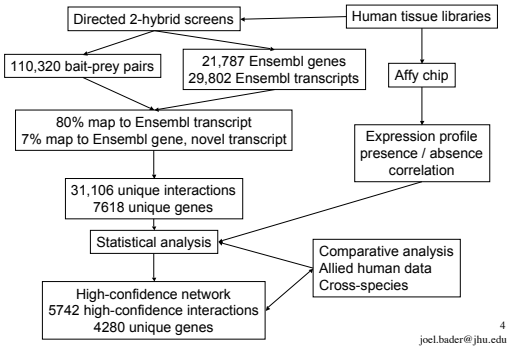
HIGH THROUGHPUT BIOLOGY

Topics

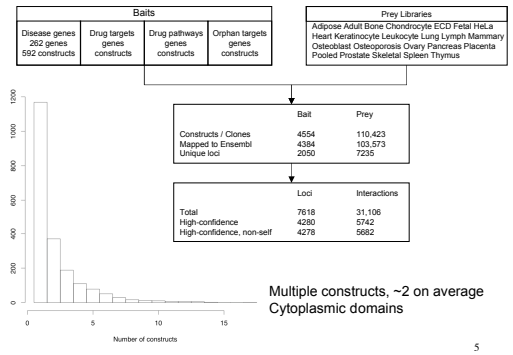
- Large-scale human protein interaction map
- Statistical framework for estimating coverage (false-negative rate) of sampled networks
- A new sequencing machine



Progress in the human network



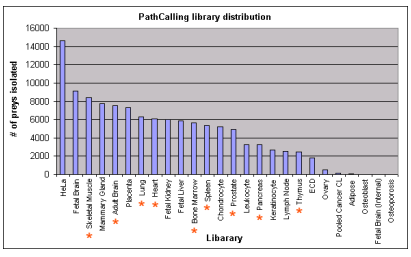
Bait design



Multiple constructs, ~2 on average
Cytoplasmic domains

Library distribution

26 PathCalling bait libraries used
- 9 of these have expression data available for same tissue
  - Affy U95A-E human chip hybridization
  - Data publicly available from NCBI GEO database
  - These 9 tissues cover 42.7% of PathCalling preys identified



* Normal tissue used for Affy U95 hybridization, public data set available from NCBI GEO database

Confidence Scores

- Framework: Logistic Regression

$$\frac{\Pr(\text{true}\,|\,\text{data})}{\Pr(\text{false}\,|\,\text{data})} = \frac{\Pr(\text{data}\,|\,\text{true})}{\Pr(\text{data}\,|\,\text{false})} \cdot \frac{\Pr(\text{true})}{\Pr(\text{false})}$$

Data = vector = (# times observed, local clustering, degree, gene region, ...)

Naive Bayes:

$$\frac{\Pr(\text{data}\,|\,\text{true})}{\Pr(\text{data}\,|\,\text{false})} \approx \prod_{\text{data type } i} \frac{\Pr(d_i\,|\,\text{true})}{\Pr(d_i\,|\,\text{false})}$$

Ignores correlation between variables

Logistic Regression:

$$\frac{\Pr(\text{data}\,|\,\text{true})}{\Pr(\text{data}\,|\,\text{false})} \approx \exp[\sum_{\text{data-type}} b_i d_i]$$

Includes correlation at the cost of a specific functional form

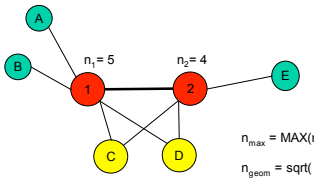
joel.bader@jhu.edu

Regression variables

Column	Type	Definition			
key1	ENSGID	Protein 1	frmpCC	INT	.. CDS / CDS
key2	ENSGID	Protein 2	frmpF1	INT	Total # of preys that are in Frame 1
allbs	STR	bootstrap (bs) or non-bootstrap (nbs) for each event	frmpF2	INT	.. F2
allexid	STR	Internal EventIDs for each event	frmpF3	INT	.. F3
allfrm1	STR	Frame designation (F1, F2, F3) for protein 1 for each event	frmpe3	INT	Total # of preys that end in 3' UTR
allfrm2	STR	Frame designation for protein 2 for each event	frmpe5	INT	.. 5' UTR
allorient	STR	Order bait-prey (bp) or prey-bait (pb) for event	frmpeC	INT	.. CDS
allrgn1	STR	Region designation (start, stop) for protein 1 for each event	frmps3	INT	Total # of preys that start in 3' UTR
allrgn2	STR	Region designation (start, stop) for protein 2 for each event	frmps5	INT	.. 5' UTR
deg1	INT	Degree of protein 1 (undirected network)	frmpsC	INT	.. CDS
deg2	INT	Degree of protein 2 (undirected network)	gala.all	STR	Alpha-gal assay results for each event
degb	INT	Out-degree of the bait (directed bait -> prey network)	gala.avg	INT	.. mean value
degp	INT	In-degree of the prey (directed bait -> prey network)	gala.max	INT	.. max value
frm13	INT	Total # baits & preys that start in 3' UTR / end in 3' UTR	gala.med	INT	.. median value
frm33	INT	.. 3' UTR / 3' UTR	gala.min	INT	.. min value
frm53	INT	.. 5' UTR / 3' UTR	gala.n	INT	.. number of available values
frm5C	INT	.. 5' UTR / CDS	galb.all	STR	Beta-gal assay results for each event
frmC1	INT	.. CDS / 3' UTR	galb.avg	INT	.. mean value
frmCC	INT	.. CDS / CDS	galb.max	INT	.. max value
frmF1	INT	Total # of baits & preys in Frame 1	galb.med	INT	.. median value
frmF2	INT	.. F2	galb.min	INT	.. min value
frmF3	INT	.. F3	galb.n	INT	.. number of available values
frmb33	INT	Total # of baits that start in 3' UTR / end in 3' UTR	nebp	INT	# of bait-prey events
frmb53	INT	.. 5' / 3'	nebs	INT	# of bootstrap events
frmb55	INT	.. 5' / 5'	nenbs	INT	# of non-bootstrap events
frmb5C	INT	.. 5' / CDS	nepb	INT	# of prey-bait events
frmbC3	INT	.. CDS / 3'	ness	INT	# of self-interactions
frmbCC	INT	.. CDS / CDS	netot	INT	Total # of events
frmbF1	INT	Total # of baits that are in Frame 1	tricnt	INT	# of topological triangles = # of bridging proteins
frmbF2	INT	.. F2	trimeanhg	INT	# triangles expected, hypergeometric distribution
frmbF3	INT	.. F3	trimeanpois	INT	# triangles expected, Poisson distribution
frmbe3	INT	Total # of baits that end in 3' UTR	trin1	INT	# of neighbors of protein 1 (ignoring self and 1-2 edge)
frmbe5	INT	.. 5' UTR	trin2	INT	# of neighbors of protein 2 (ignoring self and 1-2 edge)
frmbeC	INT	.. CDS	trinv	INT	# of other vertices in the network
frmbs3	INT	Total # of baits that start in 3' UTR	tricscorer	INT	Minus log10 of p-value for triangles, Erdos-Renyi network
frmbs5	INT	.. 5' UTR	tricscorehg	INT	Minus log10 of p-value for triangles, hypergeometric p-value
frmbsC	INT	.. CDS	tricscorepois	INT	Minus log10 of p-value for triangles, Poisson network
frmp33	INT	Total # of preys that start in 3' UTR / end in 3' UTR	xagree	BOOLEAN	Model organisms agree on true-pos / false-pos
frmp53	INT	.. 5' / 3'	xcnt	INT	# of interologs voting with value xval
frmp55	INT	.. 5' / 5'	xdir	STR	# of distinct organisms voting with value xval
frmp5C	INT	.. 5' / CDS	xval	INT	Complete record of ortholog votes
frmpC3	INT	.. CDS / 3'	xpred	DOUBLE	True-positive (1) or False-positive (0)
					Predicted confidence

joel.bader@jhu.edu

Example: Topological variables



$n_{max} = \text{MAX}(n_1, n_2)$

$n_{geom} = \text{sqrt}(\, n_1\, n_2\,)$

$n_{12} = 2$

$jac_{12} = 2 / 5$

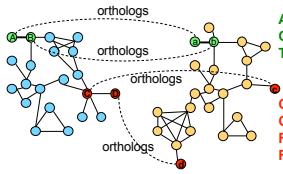
also p-value or LOD from hypergeometric distribution
(suggested by Goldberg and Roth, PNAS 2003)

joel.bader@jhu.edu

Training Set: Gold / Lead Standard

Human network

Yeast, Worm, Fly network



A / B orthologous to a / b
Connected in both networks
True positive (gold standard?)

C / D orthologous to c / d
Close together in fly network
Far apart in yeast network
False positive (lead standard?)

Compare to high-throughput only
(because later we will compare with high-confidence human sets)

joel.bader@jhu.edu

Training Set

Distance	Fly	Worm	Yeast
0	81	42	25
1	93	26	36
2	230	76	360
3	1468	295	723
4	3338	600	426
5	1687	463	154
6	352	219	28
7	66	74	5
8	15	24	2
9	1	2	0
Infinity	101	148	32

Distance 0 means In-Paralogs interact

Othologous groups determined using InParanoid algorithm
(Sonnhammer and coworkers)

joel.bader@jhu.edu

The Formula

```
glm(formula = formoct19, family = binomial, data = active, weights = weights,
    control = glm.control(trace = F))

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-1.1313  -0.2773  -0.2037  -0.1546   2.1307

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.2637     1.0004   2.263  0.02366 *
log(degb)    -0.8517     0.2929  -2.908  0.00364 **
log(degp)    -1.0731     0.4060  -2.644  0.00820 **
log(1 + frmpeC)  1.1784     0.2015   5.847 5.01e-09 ***
log(1 + frmb5C)  0.4979     0.3585   1.389  0.16490
log(degb):log(degp)  0.1740     0.1186   1.467  0.14239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 359.62  on 1759  degrees of freedom
Residual deviance: 246.30  on 1754  degrees of freedom
AIC: 148.52
```

High degree: Bad

Coding region: Good

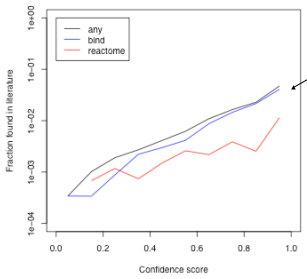
Small correction to degree

Surprise: shared neighbors are not a significant predictor (after accounting for degree)

Performance:
102 / 141 true-positives recovered (28% error)
1330 / 1619 false-positives rejected (18% error)
Entire data: ~20% high-confidence

joel.bader@jhu.edu

Overlap with Literature



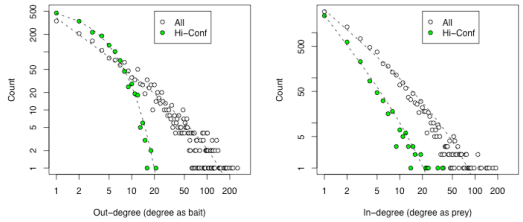
~5% of highest-confidence are found in literature

BIND 3848 tot

CRGN/JHU 5742 tot

104
2.7% of known

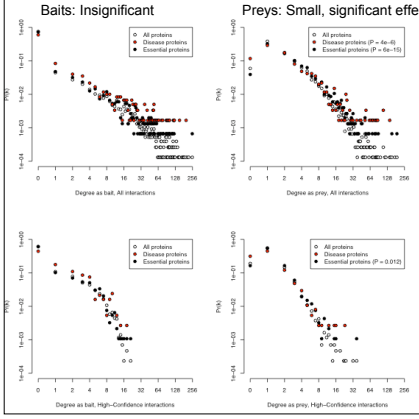
Statistical Properties



Truncated distribution for baits (limited by # of clones picked)
Long-tail distribution for prey, fit by exponentially-truncated power-law
Do disease / essential genes have more interactions?

Baits: Insignificant

Preys: Small, significant effect

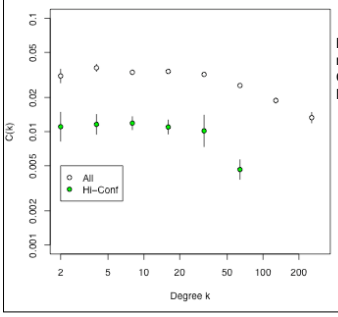


Centrality and Lethality

Disease = OMIM

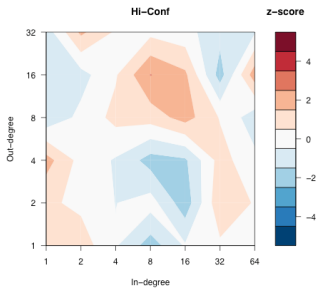
Essential = lethal deletion / RNAi phenotype for yeast, worm, or fly ortholog

Hierarchical Indicator



Barabasi/Oltvai Hierarchical network:
$C(k) \sim k^{-b}$
Not evident in this network

Assortative / Disassortative



Assortative:
Enrich connections between same degree

Disassortative:
Enrich connections between different degree
Network lore: insulates high-degree hubs

(Maslov, Sneppen: Yeast is disassortative)

Marginal (p = 0.03) evidence for disassortative

Significant motifs

Motif	Observed	Expected	Z-Score	P-value
	104	20.9±4.6	18.2	3×10^{-74}
	557	76±11	44	0
	97	24.3±7.1	10.2	1×10^{-24}
	17,333	3100±450	31.7	0
	1521	680±230	3.6	0.00016

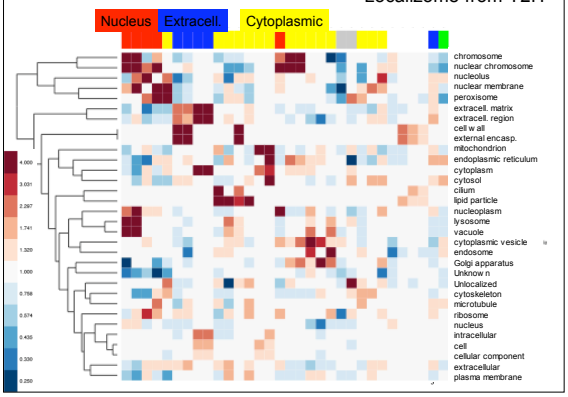
Motif	Observed	Expected	Z-Score	P-value
	1103	440±130	5.1	1.6×10^{-7}
	344	2.2±1.8	190	0
	56	2.6±3.1	17	4×10^{-65}
	504	289±28	7.6	1.5×10^{-14}
	285	61±14	16	6.4×10^{-58}
	52	9.3±6.2	6.9	2.6×10^{-12}
	30	0.8±1.0	30.	0

Enriched Domain-Domain Interactions (top 40)

desc1	desc2	pvbhi.m.8607 rich	
Hormone_recep~Family Ligand-binding domain o	Hormone_recep~Family Ligand-binding do	2.47E-15	26.587
Hormone_recep~Family Ligand-binding domain o	zf-C4~Domain Zinc finger, C4 type (two d	2.47E-15	26.587
F-box~Domain F-box domain	Skp1~Domain Skp1 family, dimerisation d	7.09E-15	159.36
F-box~Domain F-box domain	Skp1_POZ~Domain Skp1 family, tetramer	8.63E-15	157.46
bZIP_2~Family Basic region leucine zipper	bZIP_2~Family Basic region leucine zipper	3.36E-14	27.488
bZIP_1~Family bZIP transcription factor	Jun~Family Jun-like transcription factor	3.53E-13	32.693
zf-C4~Domain Zinc finger, C4 type (two domains	zf-C4~Domain Zinc finger, C4 type (two d	4.05E-13	26.145
Jun~Family Jun-like transcription factor	bZIP_2~Family Basic region leucine zipper	1.91E-12	31.566
bZIP_1~Family bZIP transcription factor	bZIP_2~Family Basic region leucine zipper	4.78E-12	24.478
bZIP_1~Family bZIP transcription factor	bZIP_1~Family bZIP transcription factor	9.36E-11	26.931
wnt~Family wnt family	Granulin~Family Granulin	1.62E-08	36.254
GTP_CDC~Family Cell division protein	GTP_CDC~Family Cell division protein	4.14E-08	868.48
Filament~Family Intermediate filament protein	Filament~Family Intermediate filament pr	5.60E-08	12.354
LSM~Domain LSM domain	LSM~Domain LSM domain	8.86E-08	115.19
TNFR_c6~Domain TNFR/NGFR cysteine-rich regio	zf-TRAF~Family TRAF-type zinc finger	1.09E-07	36.643
Peptidase_M10~Domain Matrixin	Ribosomal_L7Ae~Domain Ribosomal prote	2.10E-07	199.75
Ribosomal_L7Ae~Domain Ribosomal protein L7Ae	Peptidase_M10_N~Domain Matrix metallo	2.10E-07	199.75
Hemopexin~Repeat Hemopexin	Ribosomal_L7Ae~Domain Ribosomal prote	4.06E-07	182.35
efhand~Domain EF hand	IQ~Motif IQ calmodulin-binding motif	1.56E-06	15.925
TNFR_c6~Domain TNFR/NGFR cysteine-rich regio	MATH~Domain MATH domain	1.62E-06	28.297
Chromo~Family 'chromo' (CHRomatin Organisati	Chromo_shadow~Domain Chromo shadow	9.22E-06	977.33
Chromo_shadow~Domain Chromo shadow domai	Pre-SET~Motif Pre-SET motif	9.22E-06	977.33
Clathrin~Family Region in Clathrin and VPS	Clathrin_lg_ch~Family Clathrin light chain	1.84E-05	264.86
Clathrin_lg_ch~Family Clathrin light chain	Clathrin_propel~Repeat Clathrin propeller	1.84E-05	264.86
SH2~Domain SH2 domain	Pkinase_Tyr~Domain Protein tyrosine kina	1.98E-05	9.5489
Tubulin~Domain Tubulin/FtsZ family, GTPase dom	TB~Family TB domain	4.42E-05	161.09
TB~Family TB domain	Tubulin_C~Domain Tubulin/FtsZ family, C-	4.42E-05	161.09
zf-CXXC~Family CXXC zinc finger	SAM_2~Domain SAM domain (Sterile alph	0.00022652	130.37
SAM_1~Domain SAM domain (Sterile alpha motif	zf-CXXC~Family CXXC zinc finger	0.00024524	127.55
MH1~Domain MH1 domain	MH1~Domain MH1 domain	0.00025146	366.96
MH1~Domain MH1 domain	MH2~Family MH2 domain	0.00025146	366.96
MH2~Family MH2 domain	MH2~Family MH2 domain	0.00025146	366.96
SET~Family SET domain	Chromo_shadow~Domain Chromo shadow	0.0003488	391.05
PFK~Domain Phosphofructokinase	TB~Family TB domain	0.00042505	82.873
Chromo~Family 'chromo' (CHRomatin Organisati	Chromo~Family 'chromo' (CHRomatin Or	0.00063536	325.93
Tubulin~Domain Tubulin/FtsZ family, GTPase dom	EGF_CA~Domain Calcium binding EGF dor	0.0006812	43.633
Tubulin_C~Domain Tubulin/FtsZ family, C-termin	EGF_CA~Domain Calcium binding EGF dor	0.0006812	43.633
F-box~Domain F-box domain	Spond_N~Family Spondin_N	0.00073548	94.735
LEA_4~Family Late embryogenesis abundant prot	MAP1_LC3~Family Microtubule associated	0.00076909	196.41
Fork_head~Domain Fork head domain	Retrotrans_gag~Family Retrotransposon g	0.001002	293
SH3_1~Domain SH3 domain	SH3_1~Domain SH3 domain	0.0012041	4.9261

ader@jhu.edu

Localizome from Y2H



Disease Connections

Before multiple testing correction

After

Disease 1	Disease 2	pvbhi	pvbhi.m.70
Ataxia-telangiectasia (3)	Chondrosarcoma, extraskeletal myxoid (1)	0.00017068	0.011878
Hyperphenylalaninemia due to pterin-4a-carbinolamine	MODY, type 5 (3)	0.0003413	0.023612
Propionicacidemia, type I or pccA type (1)	Propionicacidemia, type II or pccB type (3)	0.0003413	0.023612
Breast cancer, sporadic (3)	Fletcher factor deficiency (1)	0.00034133	0.023614
Hemolytic anemia due to band 3 defect (3)	Spherocytosis-2 (3)	0.00051186	0.035205
Leukemia, acute nonlymphocytic (2)	Erythremia (1) (?)	0.00051186	0.035205
Dysfibrinogenemia, beta type (3)	Salivary gland pleomorphic adenoma (2)	0.00051186	0.035205
Diabetes mellitus, type II (3)	Chronic granulomatous disease, X-linked (3)	0.00051186	0.035205
Shah-Waardenburg syndrome, 277580 (3)	Pulmonary alveolar proteinosis, congenital, 265120 (3	0.00051192	0.035209
C8 deficiency, type II (3)	Meningioma, SIS-related (3)	0.00068236	0.046658
Glucose transport defect, blood-brain barrier (3)	Fish-eye disease (3)	0.00068236	0.046658
Neuroblastoma (3)	Prader-Willi syndrome (1) (?)	0.00068236	0.046658
Obestiy with impaired prohormone processing, 600955	Papillon-Lefevre syndrome (2)	0.00068236	0.046658
Neuropathy, hereditary sensory and autonomic, type 1	Usher syndrome, type 1D (2)	0.00068236	0.046658
Neuropathy, hereditary sensory and autonomic, type 1	Deafness, autosomal recessive 12 (2)	0.00068236	0.046658
Arts syndrome (2)	Parkinson disease, juvenile, type 2, 600116 (3)	0.00068236	0.046658
Melanoma (3)	Melanoma, 155601 (3)	0.01016	0.51073
Cardiomyopathy, familial hypertrophic (3)	Cardiomyopathy, familial hypertrophic, 2, 115195 (3)	0.020149	0.75945

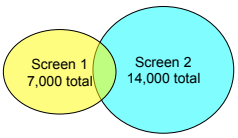
joel.bader@jhu.edu

Statistical Theory for Interactome Maps

- Unified treatment
  - Coverage of existing maps (false-negative rate, how many interactions remain to be detected)
  - Protein interaction degree distribution (power-law, scale-free, truncated, ...)
  - False-positive fraction of observed data
- Main method: Venn diagram

joel.bader@jhu.edu

The Venn diagram method



	Seen in Screen 2?	
	Yes	No
Seen in Screen 1?		
Yes	350	6,650
No	13,650	???

f1 N = 7K
f2 N = 14K
f1 f2 N = 0.35K
f1 = 0.025
f2 = 0.05
N = 280K

Problems:
Inconsistent screening methods
Screen-specific false-positives
Biases

joel.bader@jhu.edu

Statistical Framework

- Toy problem: Card counting
- A deck has k cards (hidden variable)
- Sample n cards with replacement
- Estimate Pr(k | data)
- Data
  - k_1 = # cards seen once
  - k_2 = # cards seen twice
  - k_3 = # cards seen thrice
  - ...
- Model: Y2H screen vs. normalized library
- Hailiang Huang, Prof. Bruno Jedynak

joel.bader@jhu.edu

Model Variables

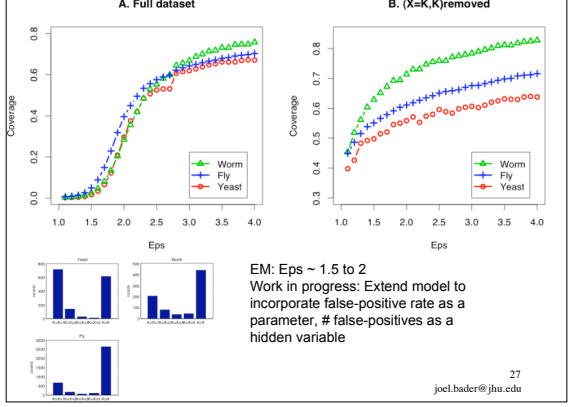
Variable	Cards	Y2H
k	Number in deck	# of interaction partners
k_i	Number seen *i* times	# preys seen *i* times
$\tilde{k} = \sum_{i>0} k_i$	Number seen at least once	# unique preys
$n = \sum_i i k_i$	Number sampled	# clones picked
$\hat{k}$	Estimate of *k*	Estimate of *k*
$\tilde{k} / \hat{k}$	Fraction of deck seen	Coverage of network

Bayesian Formula

$$\Pr(\{k_i\} \mid k,n) = \frac{k!}{\prod_i k_i!} \cdot \frac{n!}{\prod_{i>0} (k_i!)^i} \cdot k^{-n}$$

$$\Pr(k \mid \{k_i\},n) = \frac{[k!/(k-\tilde{k})!]k^{-n}\Pr(k)}{\sum_{k'}[k'!/(k'-\tilde{k})!]k^{-n}\Pr(k')}$$

Pr(*k*) = Prior estimate for *k*, power-law for scale-free network
Possible estimators:
MAP = *k* that maximizes Pr(*k*|{k_i},*n*) + 95% Confidence Interval
PME = Σ_k *k* Pr(*k*|{k_i},*n*)
Problems when $n = k_1$ (no partner observed twice): *k* = infinity
Choice for prior: Pr(*k*) = k^{-b} / zeta(b) regularizes when *b* > 1
(1) Can use EM to find best exponent *b* considering *k* as hidden variable
(2) Can look at subset of baits that detect at least 1 prey twice



EM: Eps ~ 1.5 to 2
Work in progress: Extend model to incorporate false-positive rate as a parameter, # false-positives as a hidden variable

DNA Sequencing Breakthrough

doi:10.1038/nature03959
nature

ARTICLES

Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies[1]*, Michael Egholm[1]*, William E. Altman[1], Said Attiya[1], Joel S. Bader[1], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Ksharna B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[2], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

[1]454 Life Sciences Corp., 20 Commercial Street, Branford, Connecticut 06405, USA. [2]University of California, Berkeley, California 94720, USA. [3]Laboratory of Microbiology, The Rockefeller University, New York, New York 10021, USA. [4]The Rothberg Institute for Childhood Diseases, 530 Whitfield Street, Guilford, Connecticut 06437, USA. *These authors contributed equally to this work.

The founding patent

US006274320B1

(12) **United States Patent**
Rothberg et al.

(10) **Patent No.: US 6,274,320 B1**
(45) **Date of Patent: Aug. 14, 2001**

(54) **METHOD OF SEQUENCING A NUCLEIC ACID**

(75) Inventors: **Jonathan M. Rothberg**, Guilford; **Joel S. Bader**, New Haven, both of CT (US)

(73) Assignee: **CuraGen Corporation**, New Haven, CT (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **09/398,833**

(22) Filed: **Sep. 16, 1999**

(51) **Int. Cl.**[7] **C12Q 1/68**; C12P 19/34; C12M 1/34

(52) **U.S. Cl.** **435/6**; 435/91.2; 435/287.2

(58) **Field of Search** 435/6, 287.2, 91.2

OTHER PUBLICATIONS

Baner et al., "Signal amplification of padlock probes by rolling circle replication." Nucleic Acids Research 26(22): 5073–5078 (1998).

Barshop et al., "Luminescent immobilized enzyme test systems for inorganic pyrophosphate: Assays using firefly luciferase and nicotinamide–mononucleotide adenylyl transferase or adenosine–5'–triphosphate sulfurylase." Analytical Biochemistry 197: 266–272 (1991).

Brandis et al., "Slow rate of phosphodiester bond formation accounts for the strong bias that Taq DNA polymerase shows against 2',3'–dideoxynucleotide terminators." Biochemistry 55: 2189–2200 (1990).

Bronk et al., "Combined imaging and chemical sensing using a single optical imaging fiber." Anal. Chem. 67: 2750–2757 (1996).

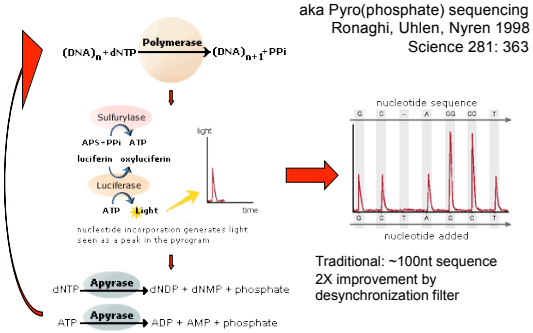
Burns et al., "An Integrated Nanoliter DNA Analysis Device." Science 282: 484–487 (1998).

Chan and Nie, "Quantum dot bioconjugates for ultrasensitive nonisotopic detection." Scient 281: 2016–2018 (1998).

Chee et al., "Accessing Genetic Information with High–Density DNA Arrays." Science 274(5287).

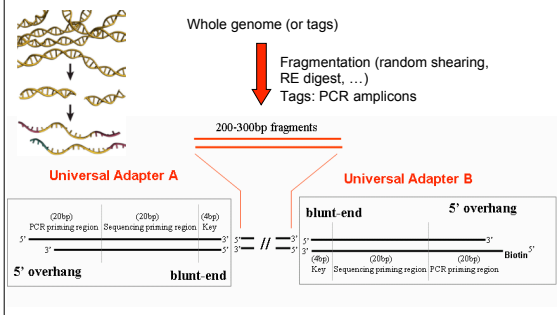
Sequencing by synthesis

aka Pyro(phosphate) sequencing
Ronaghi, Uhlen, Nyren 1998
Science 281: 363

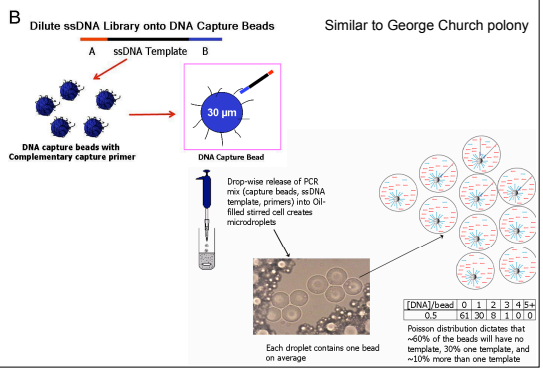


Traditional: ~100nt sequence
2X improvement by desynchronization filter

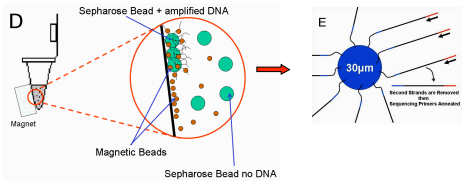
Whole genome sample prep



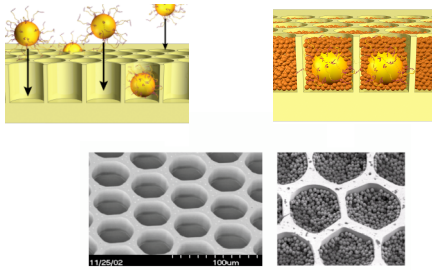
Microemulsion PCR



Recovery of clonal capture beads

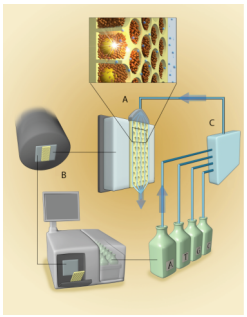


Bead loading

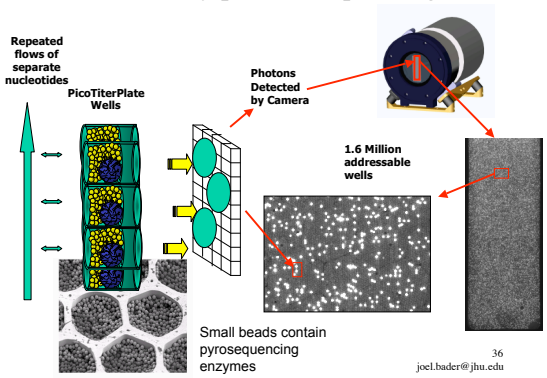


~1.6M wells per fiber optic bundle
Bundle diameter ~ 6 cm

Sequencing schematic

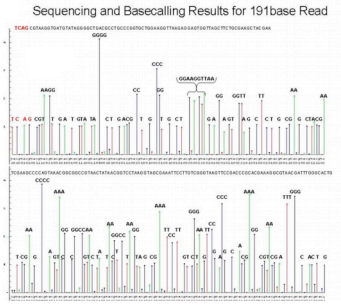


Massively parallel sequencing



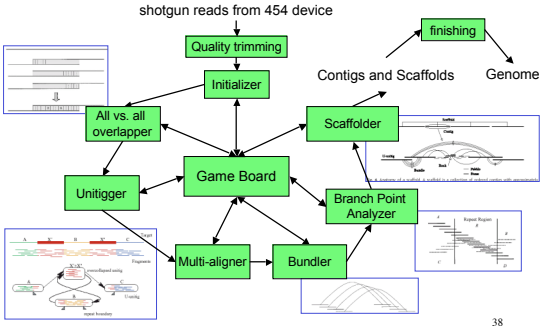
100-200 nt read length

- 95-99% accuracy per read
- 400K reads per fiber optic array
- Final assembled quality ~ Phred 30 or better for viral, bacterial genomes
- bi-directional reads (essential for vertebrate sequencing)
- Error model different from Phred: no substitutions, just homopolymer length
- Key advance: Markov model for nt +/- shift
- Luck: processive, slow lateral diffusion, functional immobilized enzymes



joel.bader@jhu.edu

NEWBLER assembler (G. Myers)



joel.bader@jhu.edu

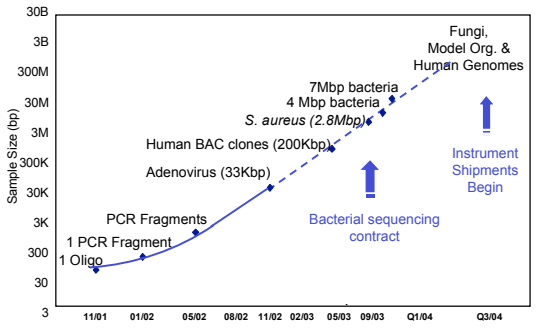
High-quality consensus sequence

Sequencing Run Summary	Adenovirus	*M. Gen.*	*S. Aureus*
Size of fiber optic slide	30x60 mm^2	60x60 mm^2	60x60 mm^2
Run Time /Number of cycles	244min/42	244 min/42	244 min/42
High Quality reads	21,904	293,066	272,602
Average read length	105 bp	115 bp	105 bp
Total number of bases sequenced	2,305,937	33,659,471	28,568,690
Individual Reads			
Reads mapped to single locations in the reference genome	20,645	221,401	255,240
Insertion error rate (Total number of overcalls/Total number of bases aligned)	2.20%	2.42%	1.31%
Deletion error rate (Total number of undercalls/Total number of bases aligned)	1.72%	2.08%	3.08%
Substitution error rate	0.09%	0.70%	0.13%
Consensus Sequence			
Number of bases aligned	2,166,283	25,743,641	27,456,492
Average oversampling	48x	37x	7x
Genome coverage	33,367 (99.7%)	568,8316 (98.1%)	2,506,618 (89.2%)
Consensus accuracy	99.994%	99.998%	99.992%
Bases overcalled	0	0.001% (6 bp)	0.003% (86 bp)
Bases undercalled	0.006% (2 bp)	0.001% (5 bp)	0.005% (124 bp)

Bacterial genomes at 99.95%+

joel.bader@jhu.edu

Moore's law improvements



joel.bader@jhu.edu

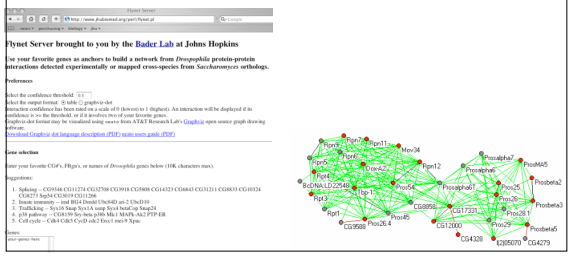
Future of 454 device

- Performance
  - ~100x higher throughput than ABI
    - 10 runs x 400K wells x 100 nt = 400 Mbp / day
    - ABI: 10 runs x 100 wells x 1000 nt = 1 Mbp / day
  - 100x improvements through feature size reductions, bi-directional reads
  - With another 10x, $1000 human
- Limitations
  - 100 to 200 nt read length
  - Errors at homonucleotide run, not at ends of sequence
- Buy one today!

joel.bader@jhu.edu

Web services and prototypes

- www.jhubiomed.org/flynet
- coming soon: HistoneDB, Systematic MUTation DB
- All software freely available, open source license (Perl, R, C, Apache/mod_perl on Mac / FreeBSD server)



Acknowledgments

- Bader lab
  - Dr. Ping Ye, Dr. Shan Guan, Dr. Limin (Angela) Liu
  - Hailiang Huang, Yan Qi, Alex Martaens
- JHU: Jef Boeke lab, Forrest Spencer lab (SL), Akhilesh Pandey
- CuraGen / 454: Jonathan Rothberg, John Chant (Genentech), Amit Chaudhuri, Shu-Xia Li, Carol Pena, Jim Knight, Martin Leach, John Simpson
- HMS: Norbert Perrimon, Drosophila RNAi Center, Fritz Roth
- Funding: NIGMS, NCRR (Roadmap Center for Networks & Pathways of Lysine Modifiction, aka Special-K Center), Whitaker Foundation

www.jhubiomed.org

44
joel.bader@jhu.edu