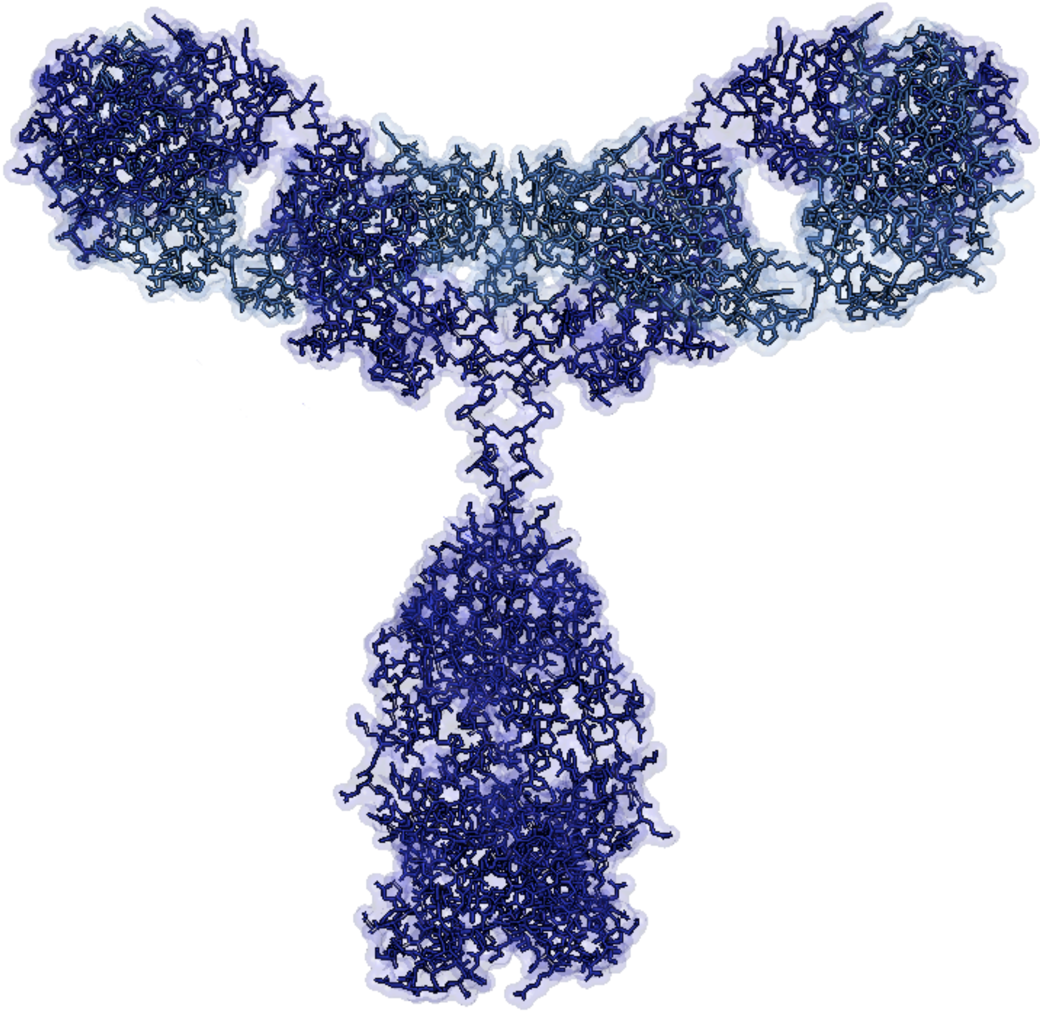
Antibodies: Nature's Solution to Molecular Recognition

Antibodies target abnormal material for immune neutralization/clearance

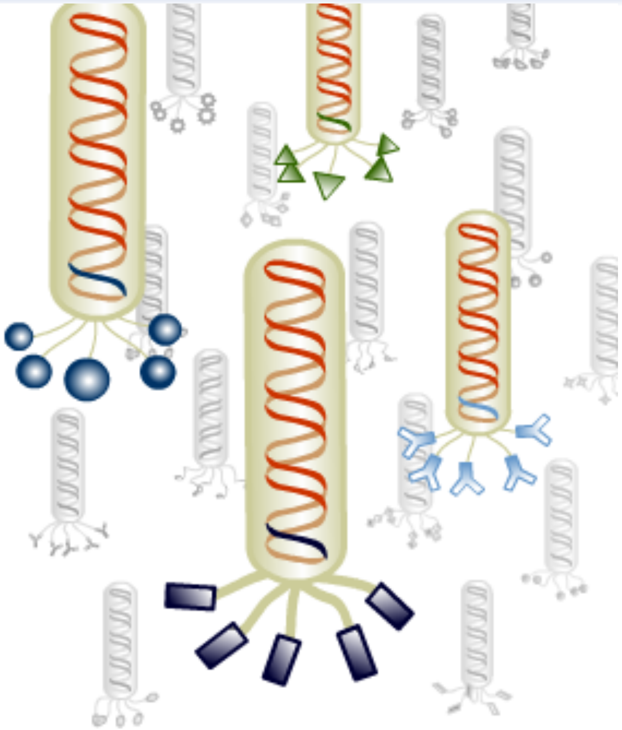
Up to 1 billion unique antigen recognition surfaces in a healthy adult

Engineered antibodies emerging as dominant new class of next generation therapeutics



Ultra-High throughput screening: Phage display

Phage display libraries: engineered virus populations that can display human antibodies on their surface coats



Enables human antibody selection

- □ All human germline families in Pfizer-lib

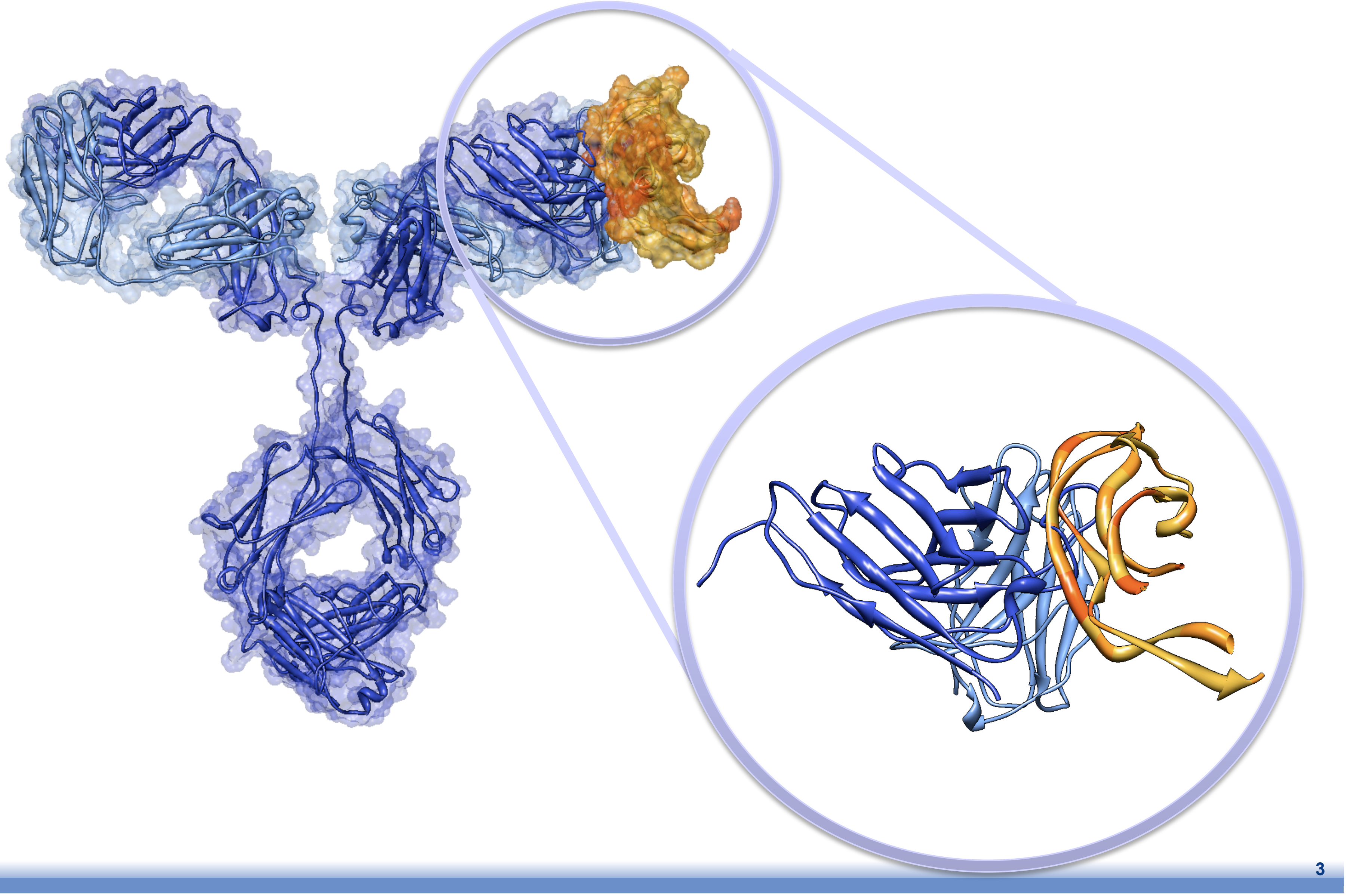
Enables immune system pooling

- □ 654 human donors combined in Pfizer-lib

Enables ultra-high throughput screening

- □ Forty billion transformants in Pfizer-lib

The single-chain Fv antibody fragment



Pfizer donor-derived IgM scFv library

Materials

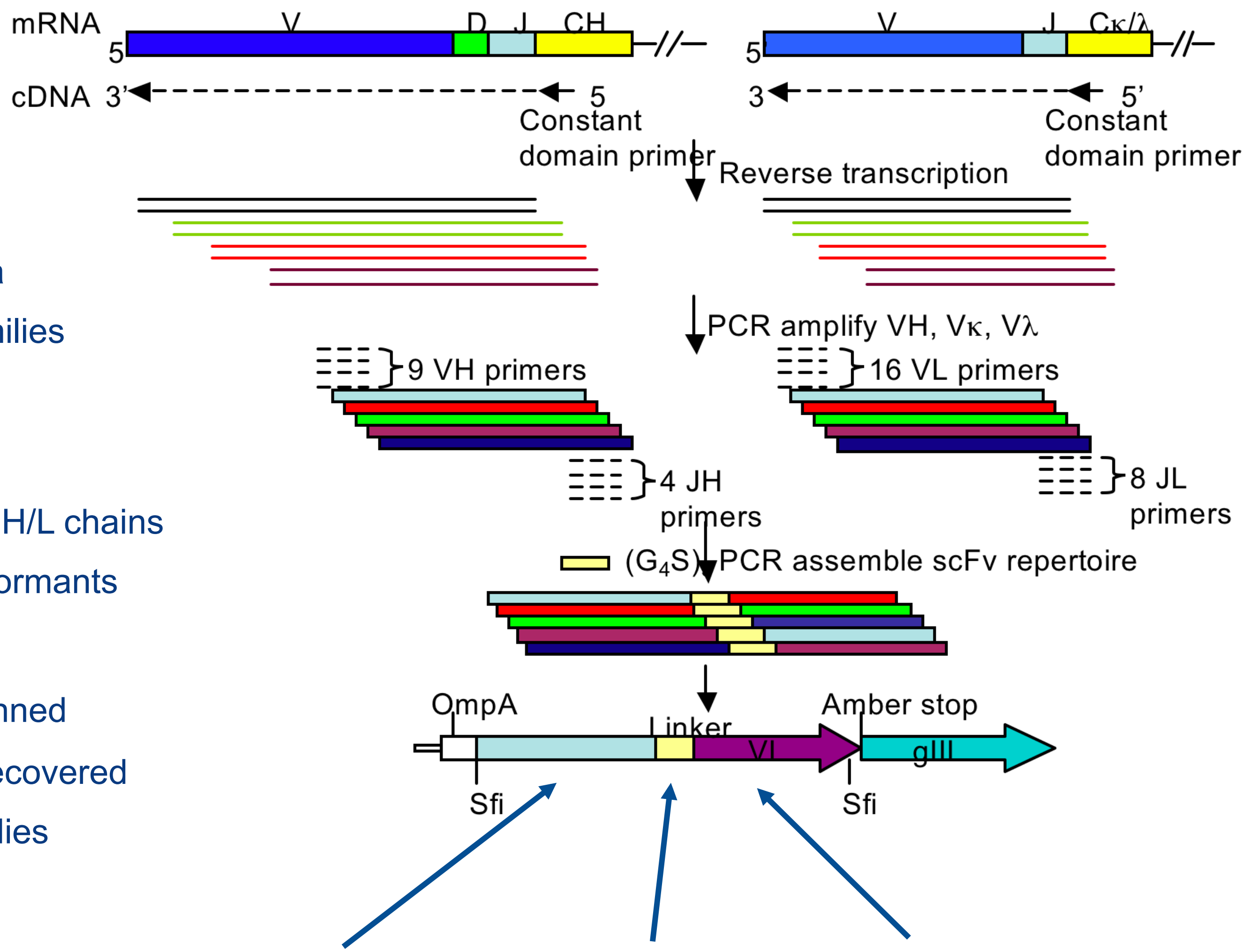
- 654 human donors
- IgM Vh amplification
- Naïve, memory, plasma
- All human germline families

Construction

- scFv format
- Random assortment of H/L chains
- 3.1+/-0.7x10E10 transformants

Panning

- 16 diverse antigens panned
- >100,000 sequences recovered
- >20,000 unique antibodies

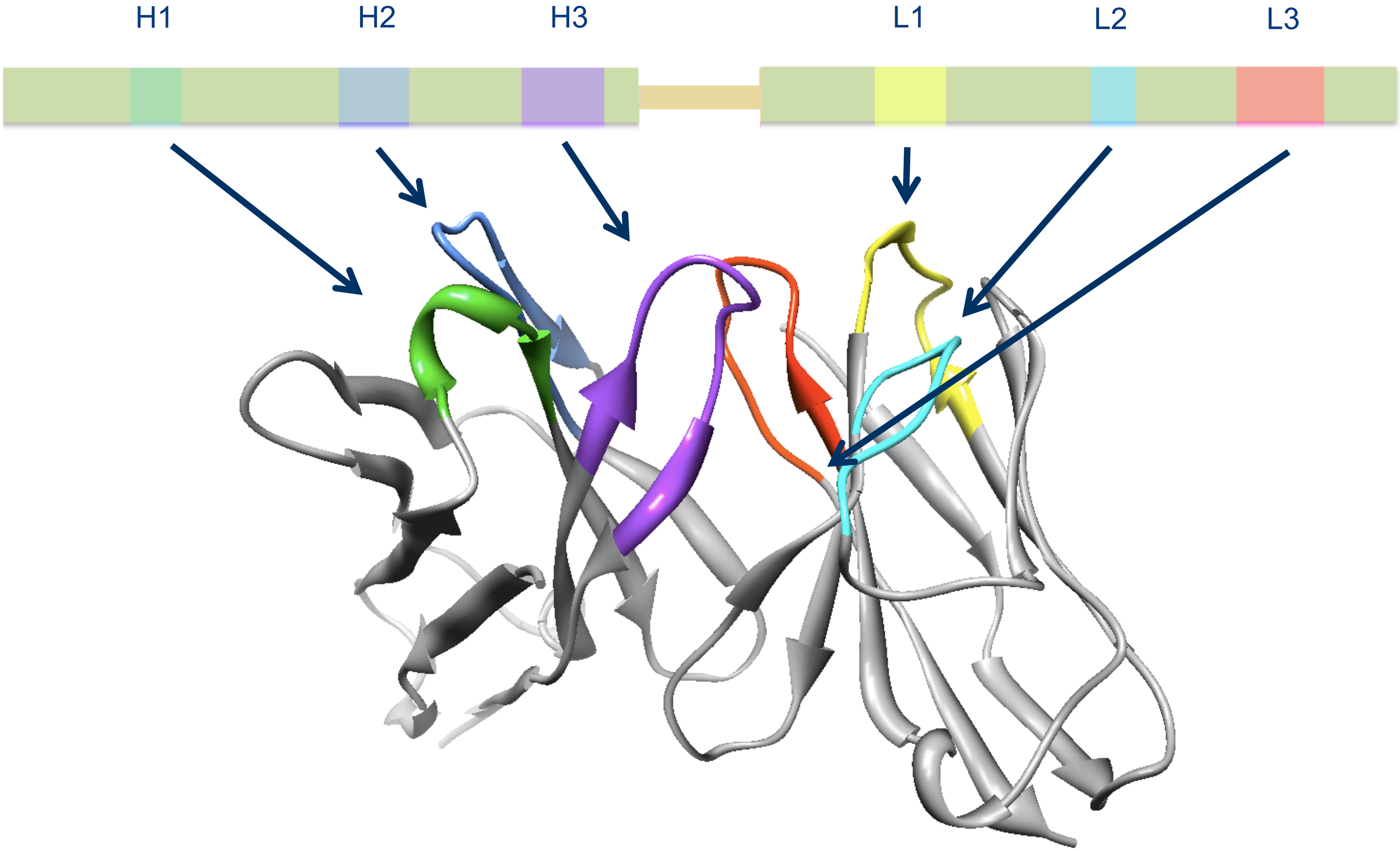


350-400bp

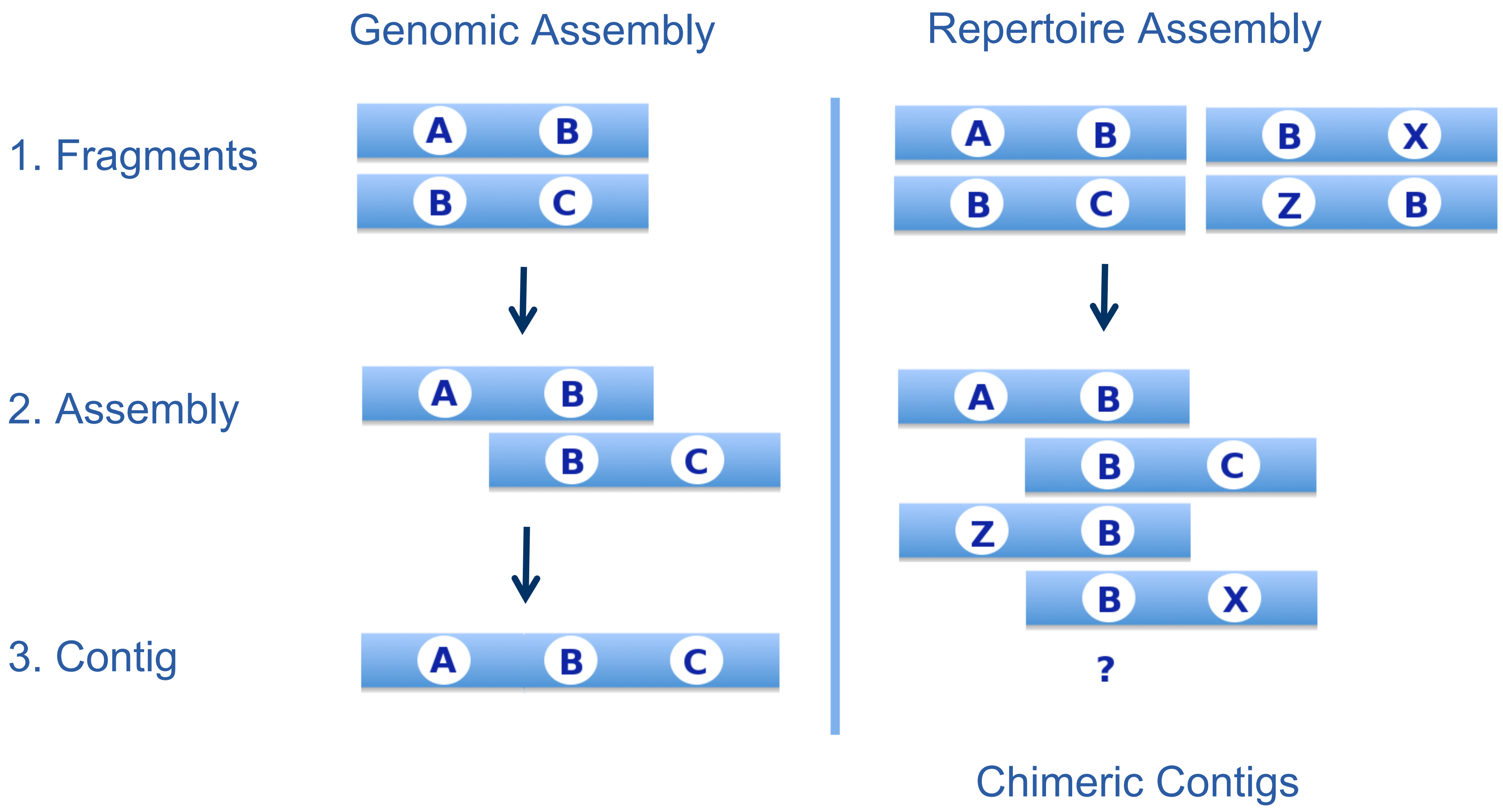
45bp

300-350bp

The scFv binding surface is discontinuous over 900bp



Antibody Diversity Prohibits Assembly



Genomic Assembly

Repertoire Assembly

1. Fragments

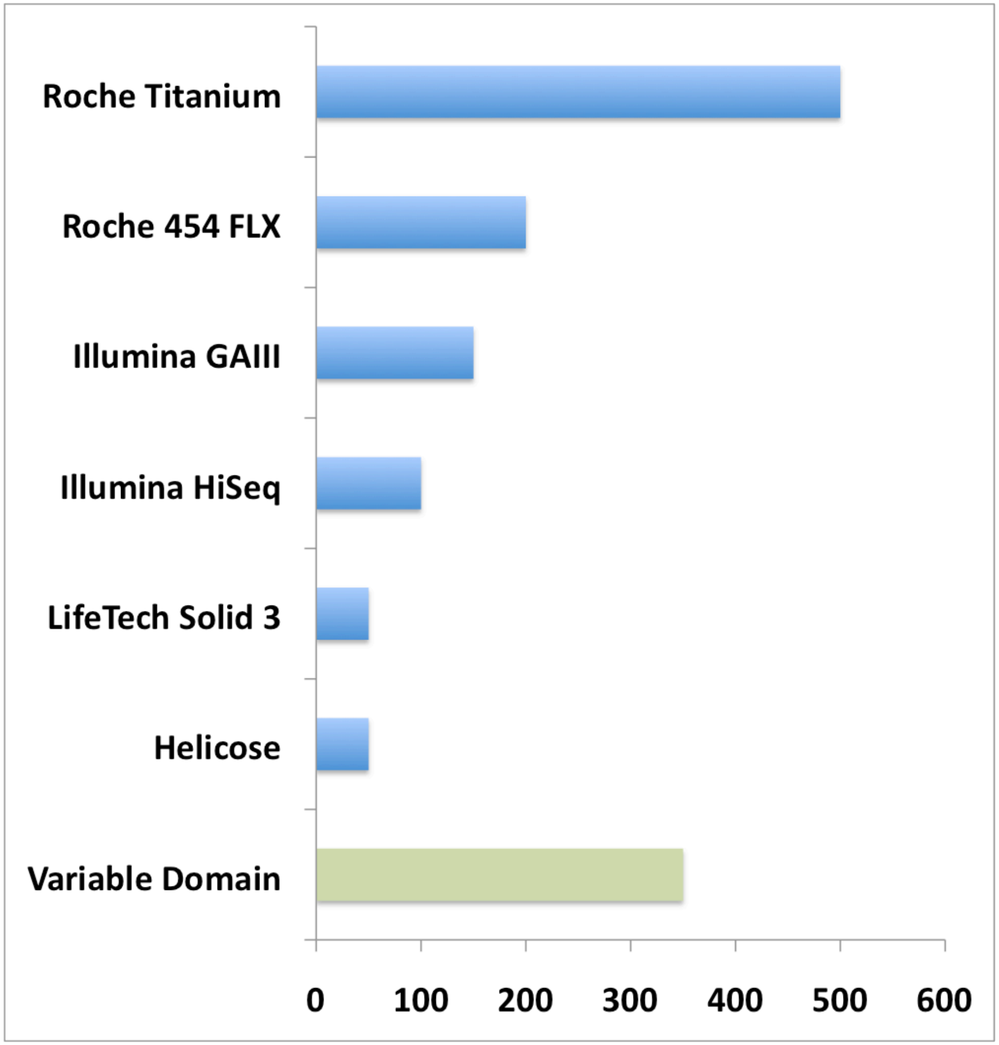
2. Assembly

3. Contig

Chimeric Contigs

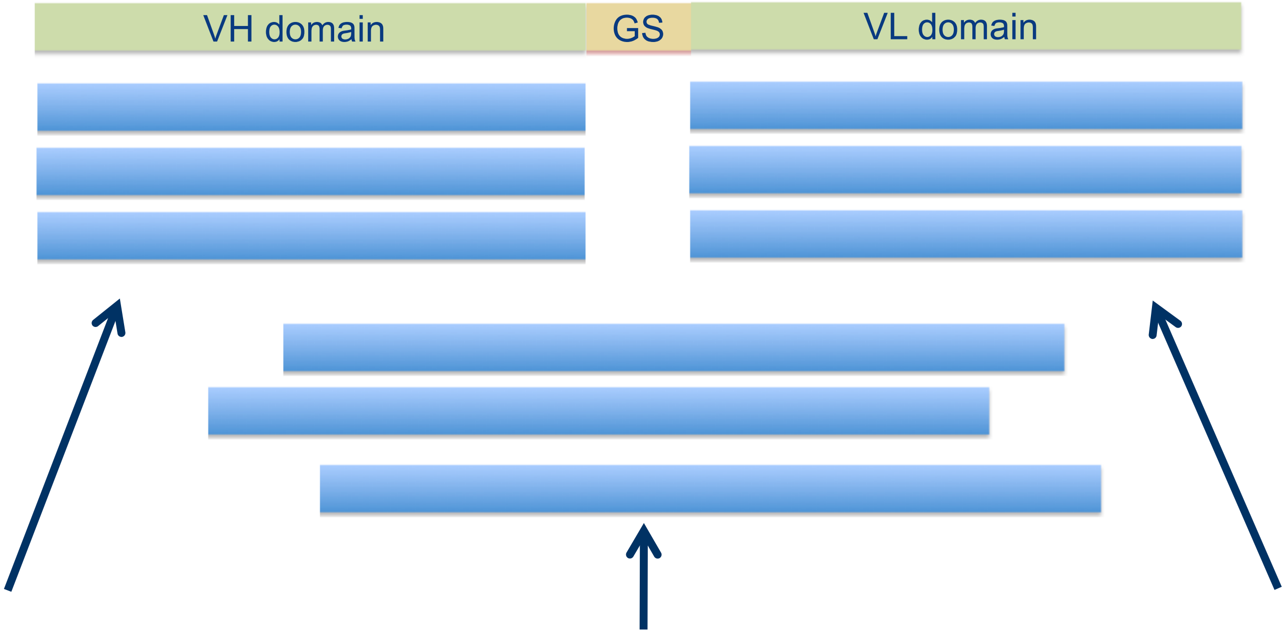
Read Lengths of Available Next Generation Sequencers

Instrument	Reads [10^6]	Read length
Roche 454 GS FLX	1-2	250-500bp
Illumina GA III	138-168	150 (2x75)bp
Illumina HiSeq 2000	<1000	2x100bp
Life Tech SOLiD 3	400	25-50bp
Helicos HeliScope	400	25-50bp



Minimum Read Length requirements for diversity estimate

scFv Insert Architecture



VH amplicons

Evaluate VH Diversity

350-400bp

RCA shotgun reads

Evaluate H/L Chain Pairing

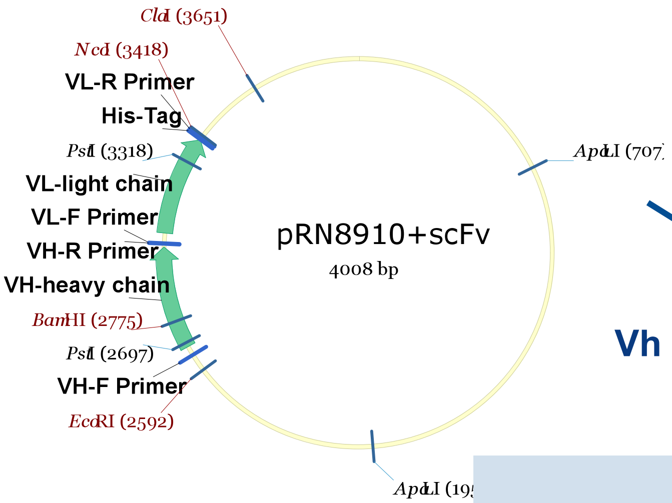
>400bp

VL amplicons

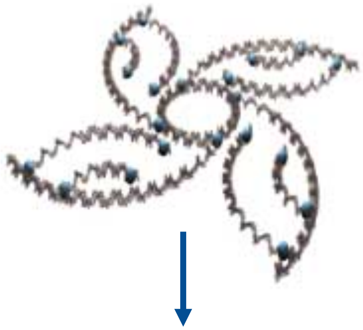
Evaluate VL Diversity

350bp

Assessment of a Phage Display Library Diversity by 454 Sequencing - Sample Preparation Strategies



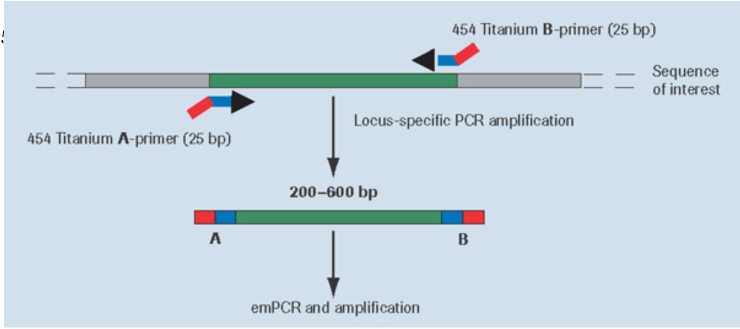
Templiphi Rolling Circle Amplification



Random shearing

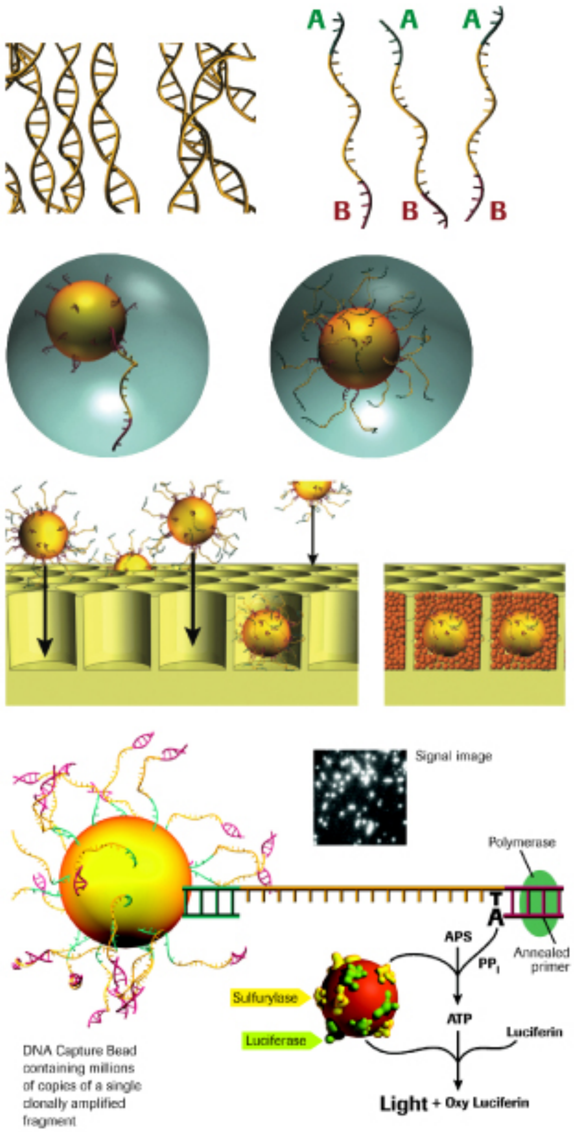
454 Library construction

Vh / Vl PCR with 454-Flap PCR primers


454 Titanium B-primer (25 bp)
Sequence of interest
454 Titanium A-primer (25 bp)
Locus-specific PCR amplification
200–600 bp
A
B
emPCR and amplification

454 emulsion-PCR and Sequencing

Sequences Obtained



Raw Reads

- 554,310 amplicon reads
- 923,875 RCA shotgun reads

High Quality Reads

- 96,303 Full VH in-frame reads
- 98,946 Full VK/L in-frame reads

How to meaningfully quantify diversity

A definition of diversity that measures unique binding surfaces

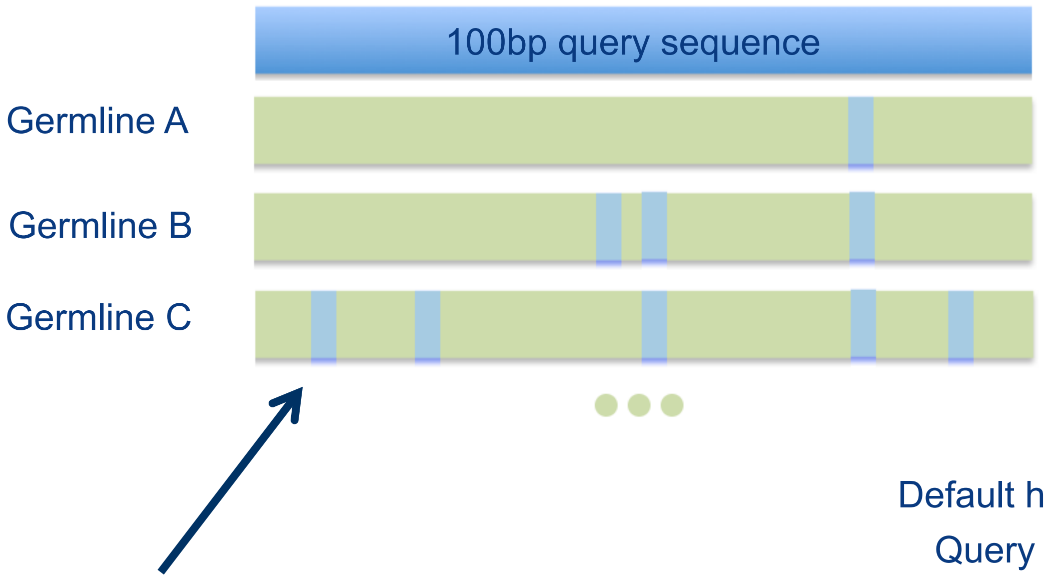
- Single mutations in CDRs probably won’t be enough to change the recognition potential
- Mutations outside the CDRs & Vernier zones are much less likely to fundamentally alter the binding profile
- Substitution errors, while rare, do occur
- Silent mutations have no effect on binding at all

- **Solution: Capture recapture**
- **Non-redundant CDR amino acid diversity definition**

How to reliably classify germline origins?

$$S_i = 1 - \sum_{g=1}^{G} \frac{P((\lambda - \delta_{ig}),(\mu_g - \delta_{ig}))}{P(\lambda,\mu_g)} \Big| g \neq i$$

Asks “what are the odds that mutations in very specific positions would cause me to erroneously classify this sequence?”



L=100,M=1; Default hypothesis

L=100,M=3,D=2; P= (98/161700) = 6.1e-4

L=100,M=5,D=4; P= (96/3921225) = 2.4e-5

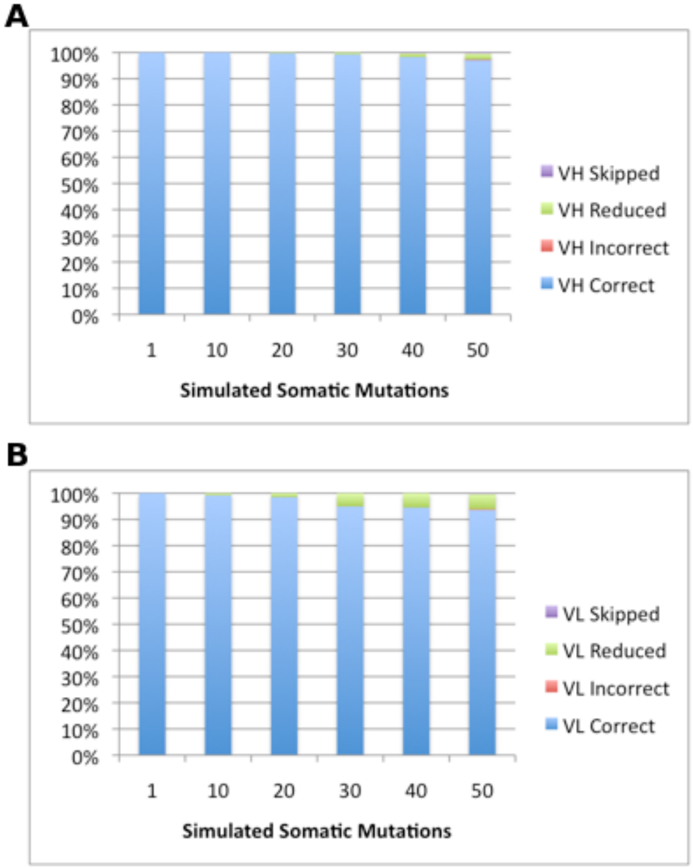
Differences between query and germline

Default hypothesis S= 1 - 6.1e-4 - 2.4e-5 = 99.94%

Query is Germline A with 99.94% confidence

Validation of probabilistic germline classification

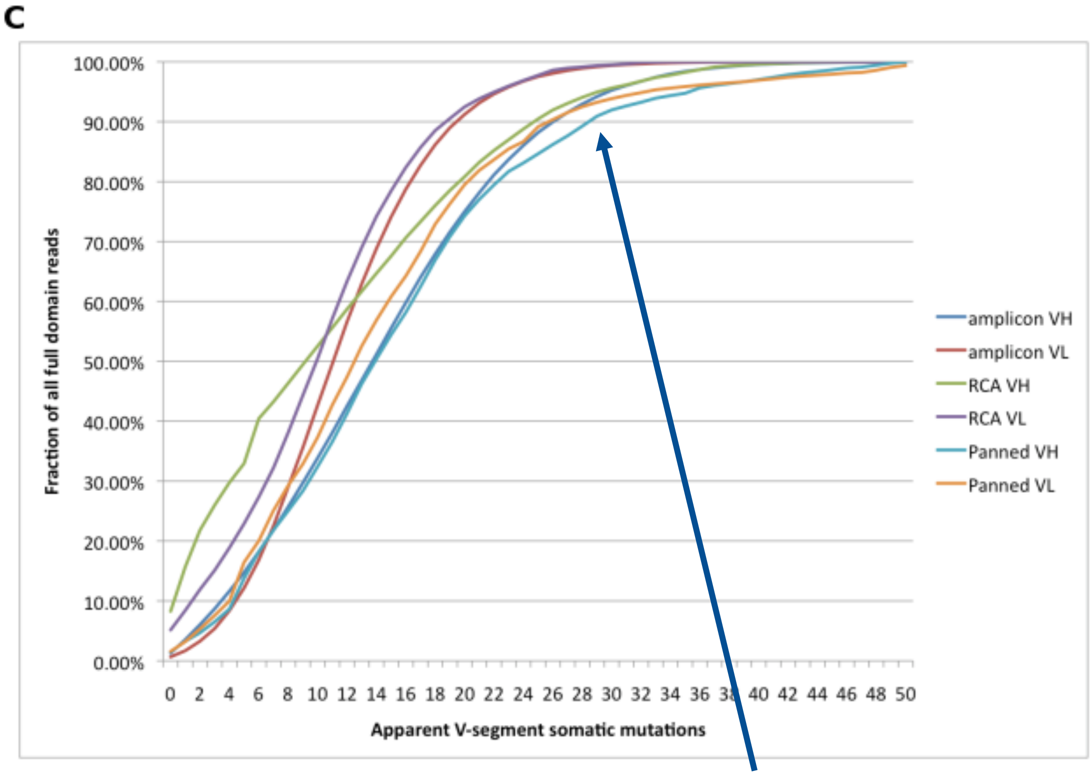
Germline sequences
randomly mutated and reclassified



8 errors in 250,000 classifications

All errors occurred when over 40
simulated mutations had been applied

Somatic mutation rate
in actual sequence data



95% of actual sequences recovered had less than 30
Mutations from closest germline framework

The Challenge of CDR Recognition

Multiple revised numbering systems

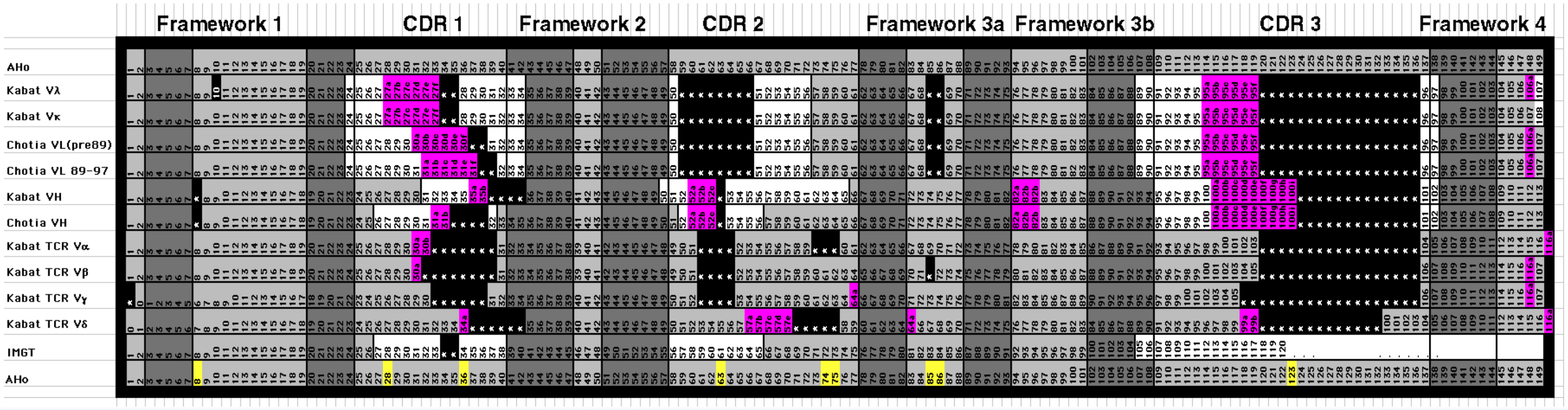
- Kabat
- Chotia
- Aho

Difficulty applying numbering systems

- 10% Kabat sequences mislabeled by own numbering system
- Rosetta Antibody Modeler fails to identify all CDRs in 30% of cases

Cause of problem?

- Length diversity
- somatic hypermutation

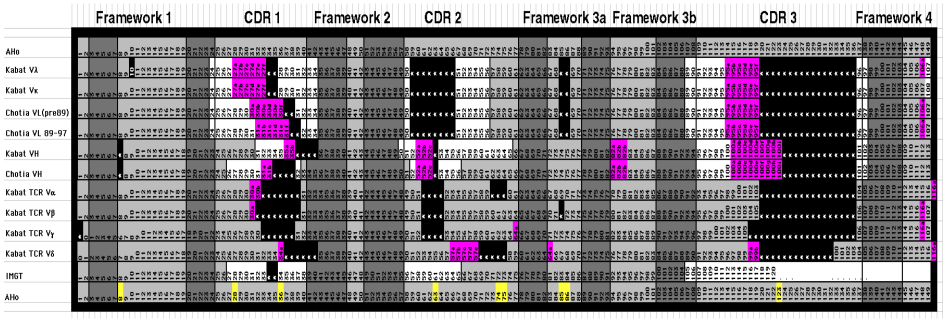

Framework 1
CDR 1
Framework 2
CDR 2
Framework 3a
Framework 3b
CDR 3
Framework 4
AHo
Kabat Vλ
Kabat Vκ
Chotia VL(pre89)
Chotia VL 89-97
Kabat VH
Chotia VH
Kabat TCR Vα
Kabat TCR Vβ
Kabat TCR Vγ
Kabat TCR Vδ
IMGT
AHo

Kabat-Labeled HMM Construction

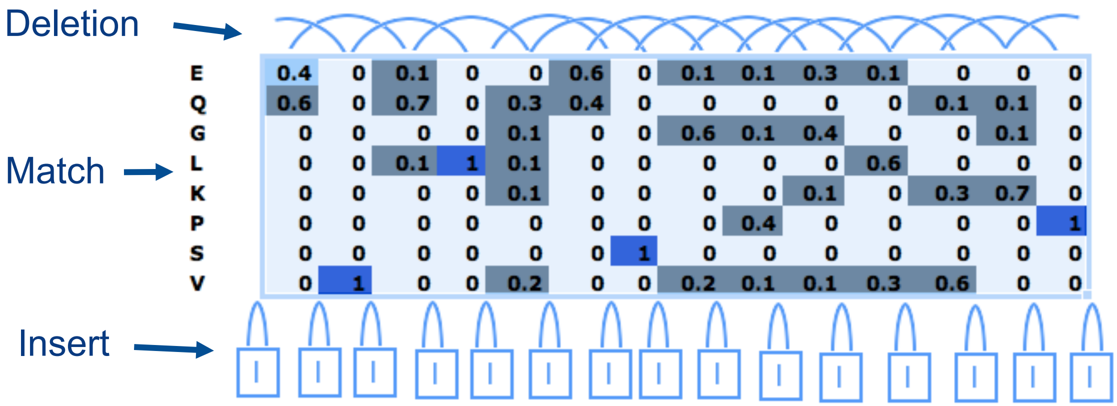
Amino Acid Alignment

HV1	Q	V	Q	L	G	Q	S	E	V	E	V	K	K	P
HV2	Q	V	V	L	K	E	S	G	P	V	L	V	K	P
HV3	E	V	E	L	V	E	S	V	E	K	E	Q	Q	P
HV4	E	V	Q	L	V	E	S	G	G	G	L	V	K	P
HV5	Q	V	Q	L	Q	E	S	G	P	G	L	V	K	P
HV6	E	V	Q	L	L	Q	S	V	V	E	V	K	G	P
HV7	Q	V	Q	L	Q	Q	S	G	P	G	L	V	K	P
Consensus	Q	V	Q	L	Q	E	S	G	P	G	L	V	K	P

Kabat numbering


Framework 1
CDR 1
Framework 2
CDR 2
Framework 3a
Framework 3b
CDR 3
Framework 4
AHo
Kabat Vλ
Kabat Vκ
Chotia VL(pre89)
Chotia VL 89-97
Kabat VH
Chotia VH
Kabat TCR Vα
Kabat TCR Vβ
Kabat TCR Vγ
Kabat TCR Vδ
IMGT
AHo

Hidden Markov Model (HMMER)



E	0.4	0	0.1	0	0	0.6	0	0.1	0.1	0.3	0.1	0	0	0
Q	0.6	0	0.7	0	0.3	0.4	0	0	0	0	0	0.1	0.1	0
G	0	0	0	0	0.1	0	0	0.6	0.1	0.4	0	0	0.1	0
L	0	0	0.1	1	0.1	0	0	0	0	0	0.6	0	0	0
K	0	0	0	0	0.1	0	0	0	0	0.1	0	0.3	0.7	0
P	0	0	0	0	0	0	0	0	0.4	0	0	0	0	1
S	0	0	0	0	0	0	1	0	0	0	0	0	0	0
V	0	1	0	0	0.2	0	0	0.2	0.1	0.1	0.3	0.6	0	0

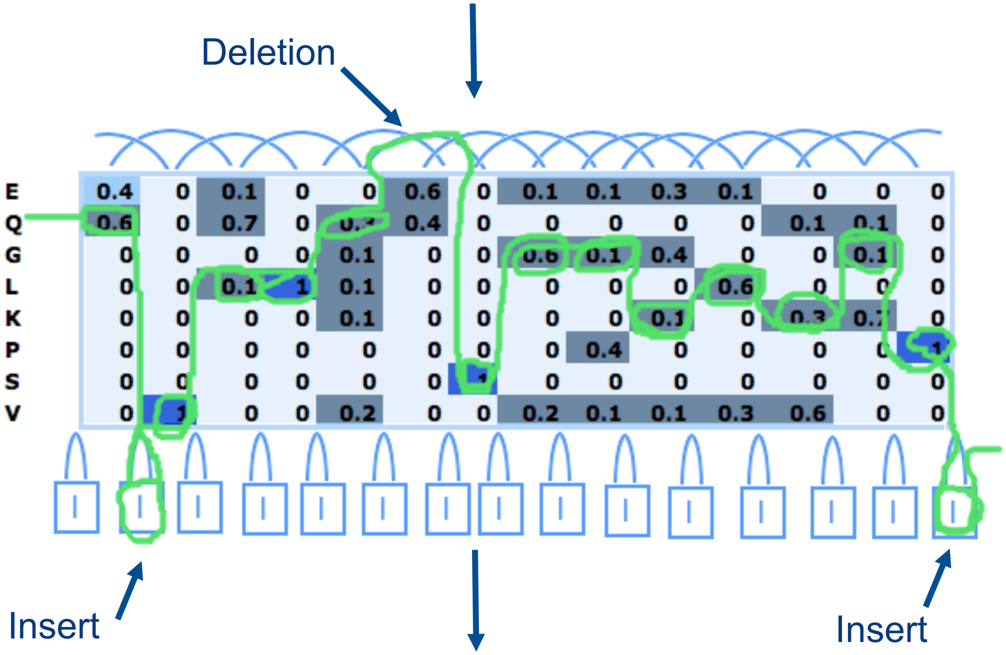
How to identify and align diverse antibody sequences?

Optimal path through HMM determined probabilistically

Probability of resulting sequence compared to probability of a random sequence (e-value)

Sequences with low e-values identified as bearing ig-like content

```
Query    QMVLLQSGGKLKGPNY
```



E	0.4	0	0.1	0	0	0.6	0	0.1	0.1	0.3	0.1	0	0	0
Q	0.6	0	0.7	0	[illegible]	0.4	0	0	0	0	0	0.1	0.1	0
G	0	0	0	0	0.1	0	0	0.6	0.1	0.4	0	0	0.1	0
L	0	0	0.1	1	0.1	0	0	0	0	0	0.6	0	0	0
K	0	0	0	0	0.1	0	0	0	0	0.1	0	0.3	0.7	0
P	0	0	0	0	0	0	0	0	0.4	0	0	0	0	1
S	0	0	0	0	0	0	1	0	0	0	0	0	0	0
V	0	1	0	0	0.2	0	0	0.2	0.1	0.1	0.3	0.6	0	0

```
HMM Path Becomes an alignment:
Consensus          Q-VQLQESGPGLVKP--
Query              QMVLLQ-SGGKLKGPNY
```

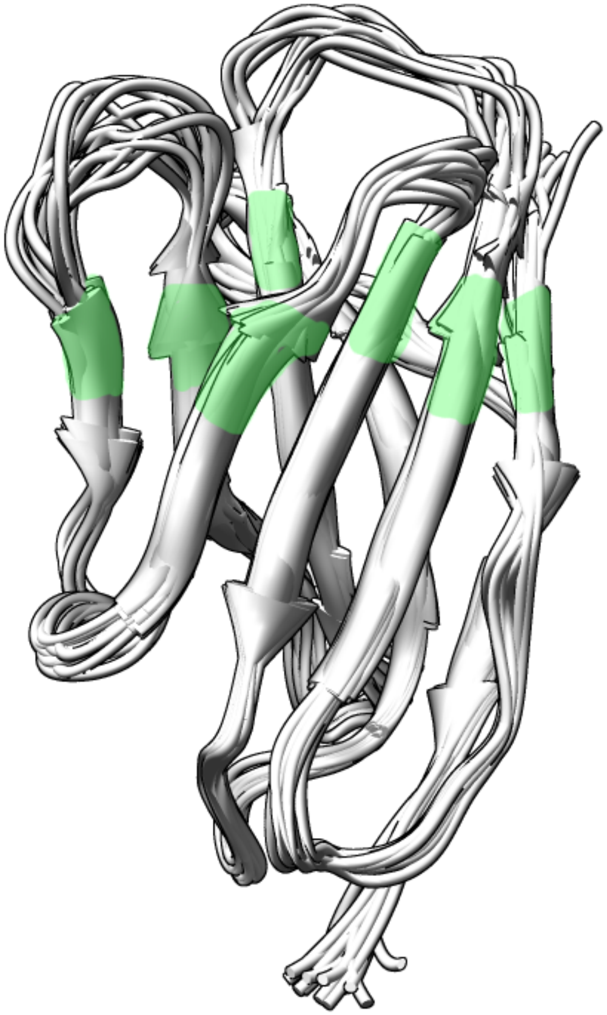
Validation of CDR Recognition accuracy

HMM CDR recognition was evaluated structurally

- 779 reference structures were structurally superposed
- Sequences of references structures were extracted
- Reference structure sequences were aligned to HMM
- Predicted boundary positions were compared to structure

HMM CDR recognition was highly accurate

- 99.93% boundary recognition accuracy



Amplicons Establish Non-Redundant CDR Diversity

Raw Reads

- 554,310 amplicon reads
- 923,875 RCA shotgun reads

High Quality Reads

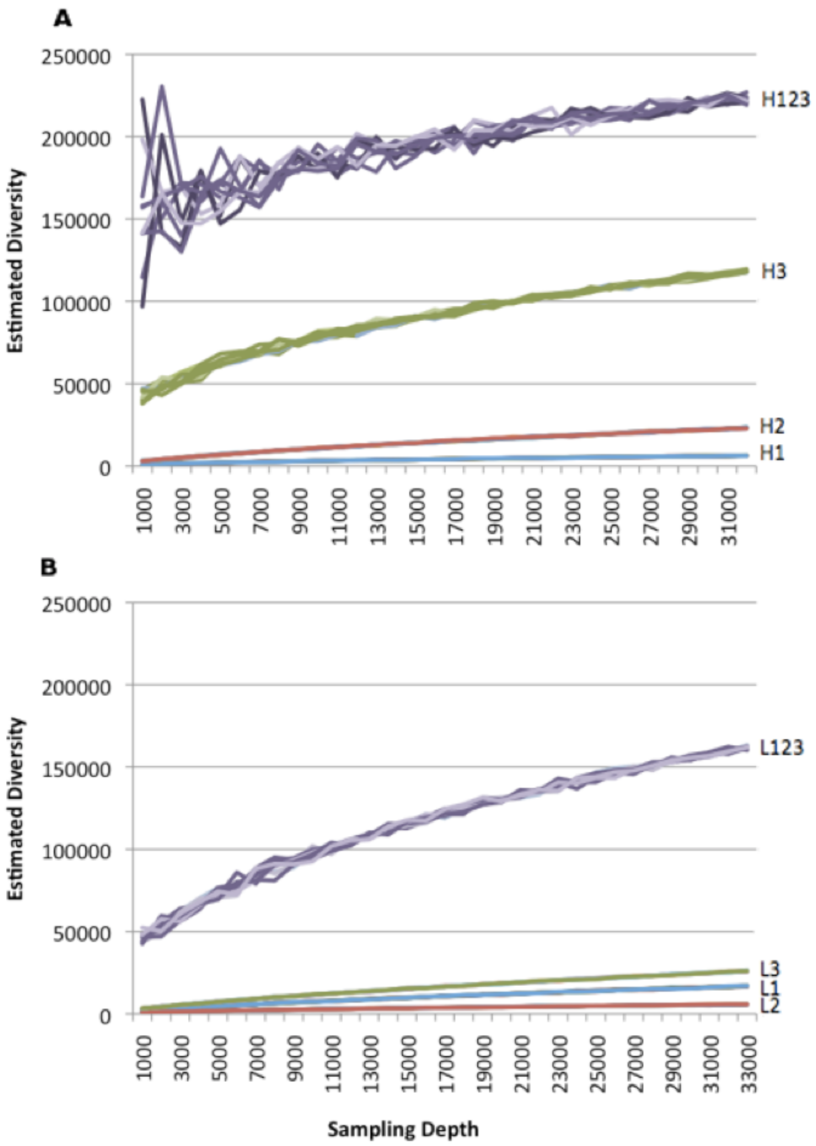
- 96,303 Full VH in-frame reads
- 98,946 Full VK/L in-frame reads

NR-Vh CDR diversity

- CDR-H1: 10E2
- CDR-H2: 10E4
- CDR-H3: 10E5
- Total Vh CDRs: 2.2+/-0.2 * 10E5

NR-Vl/k CDR diversity

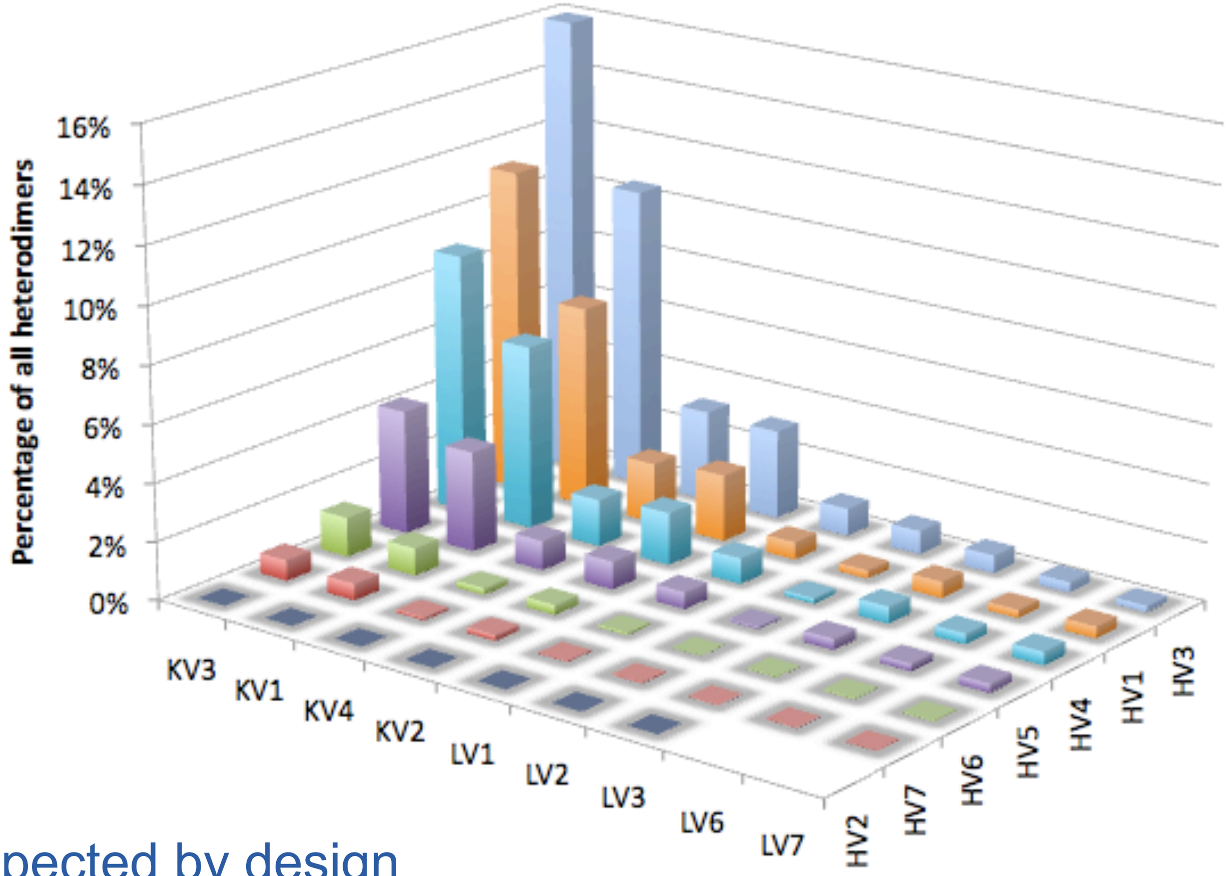
- CDR-L1: 10E3
- CDR-L2: 10E2
- CDR-L3: 10E4
- Total Vh CDRs: 1.6+/-0.8 * 10E5



RCA Shotgun Confirms H/L Random Assortment

Raw reads

- 923,875



95.6% of GS-linker expected by design

Full length clones dominate library

Random H/L assortment

Total paratope diversity estimates possible

Effects: Estimating diversity of donor derived library

- 1.5 million variable domain sequences obtained by 454 Roche Titanium chemistry pyrosequencing
- Developed novel application of Kabat column-labeled profile Hidden Markov Models (HMM)
- Used capture-recapture estimates with rarefaction to estimate total functional paratope diversity

- Forty billion distinct binding surfaces ($4 \mathrm{x} 10^{10}$)

Glanville, Zhai, Berka et al. 2009, *PNAS*

Future Applications: Immune Surveillance

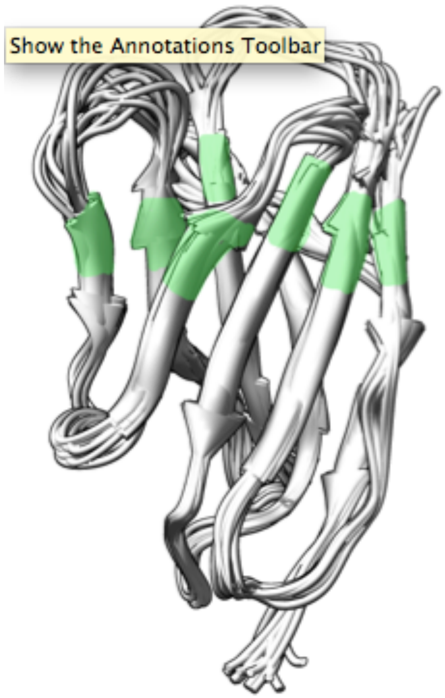
Optimized phage display library design

- QC products at multiple stages of library assembly
- Modify library designs to optimize functional diversity

Adaptive immune repertoire surveillance

- Patient stratification
- Autoimmunity
- Pathogen response
- Subunit vaccine optimization
- Powerful biomarker for primary research

Sequence Analysis Summary: Beyond Assembly

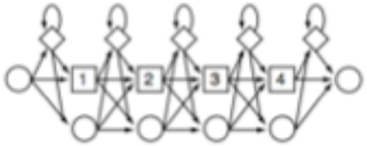


Obstacles:

- □ Repertoire still too large for achieving coverage
- □ Somatic hypermutation prohibits assembly
- □ CDR recognition not trivial
- □ Indel errors during homopolymeric stretches
- □ Correlation of CDR mutations required

Solutions:

- □ Assembly-free sequencing
- □ Hidden Markov Model References
- □Titanium Chemistry long reads
- □ Shotgun RCA for H/L independence
- □ Capture-recapture of NR translated CDRs for diversity



LIMS: Conclusions

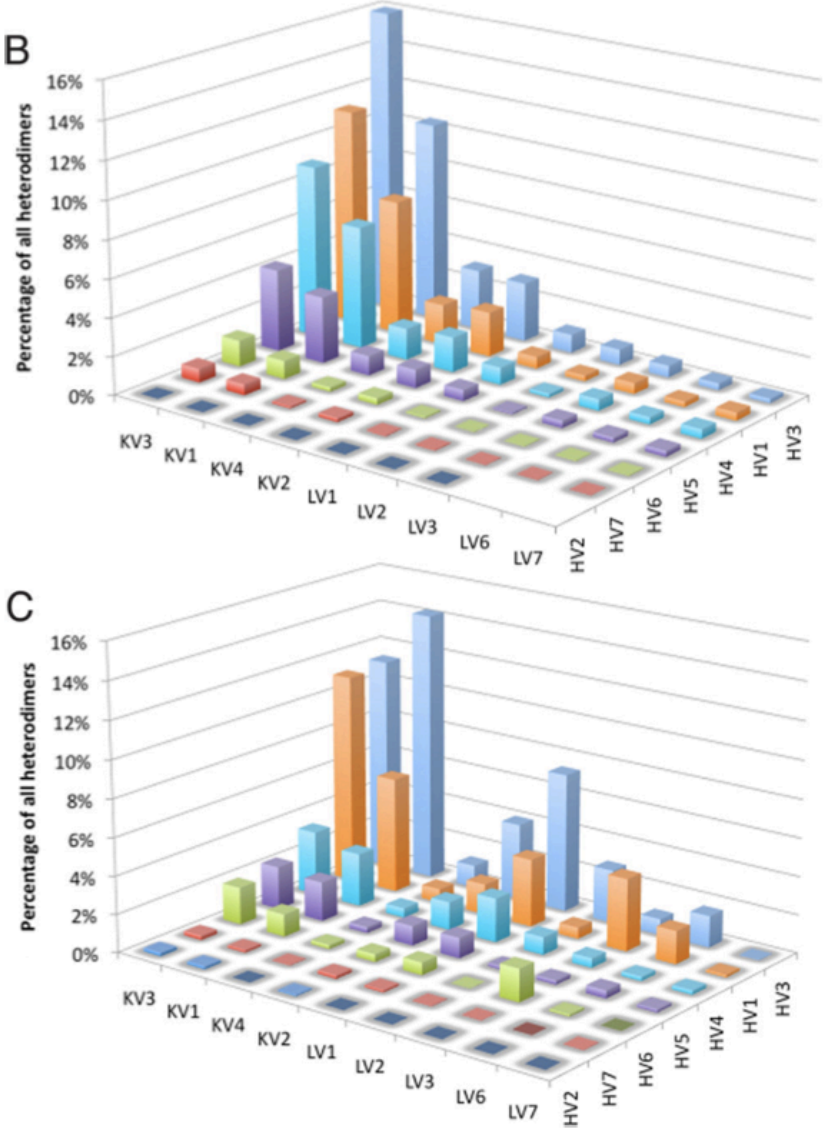
454 Sequencing Environment enabled development

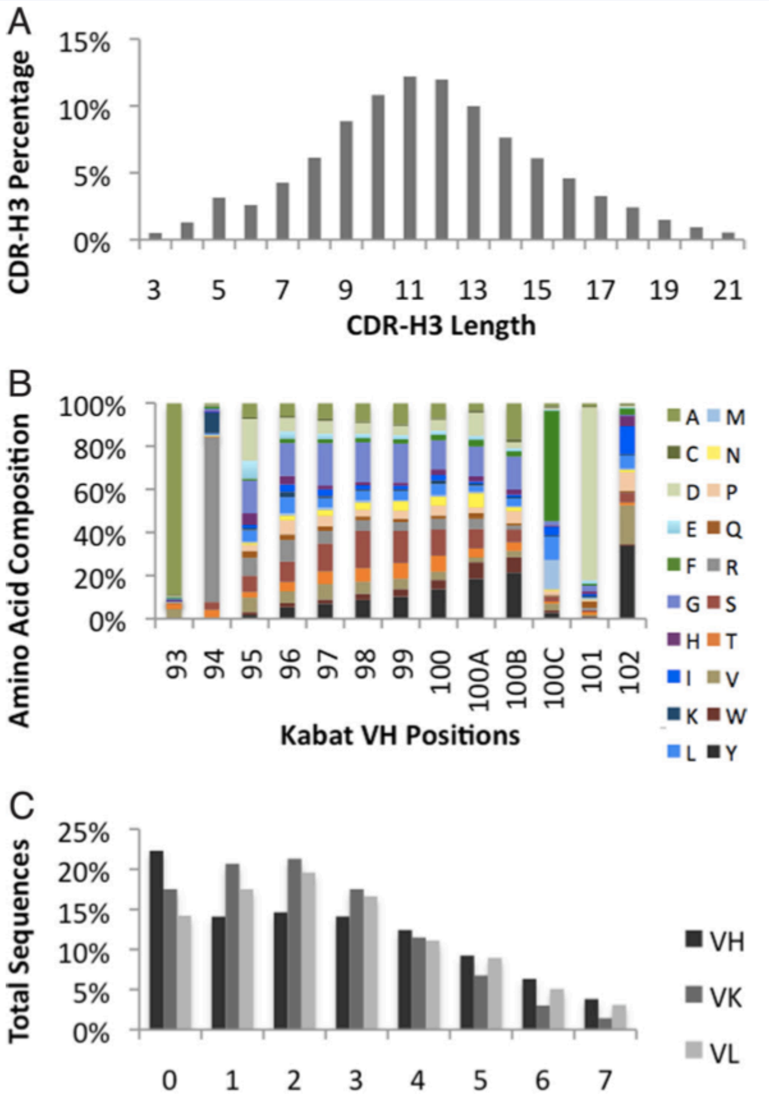
- Linux operating system on instrument and Titanium cluster
- Open data repository and well-documented data structure

WikiLIMS enabled rapid data interface development

- Direct filesystem access to 454 Sequencing files
- Modular embedded database views from external sources
- Efficient data & analysis display
- Coast-to-coast data sharing
- Backup monitoring

Diversity of donor-derived library





Glanville, Zhai, Berka et al. 2009, *PNAS*