The evolution of MIP technology

Yiyang

2013-3-15

all possible combination of primers

n PCR primer pairs,
n target sequences,
$(2n^2+n)$ possible
pairwise primer
combinations.

cross-reaction
among primer
pairs!
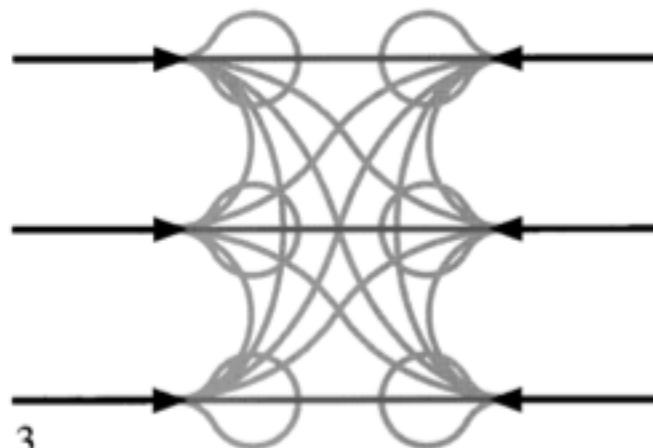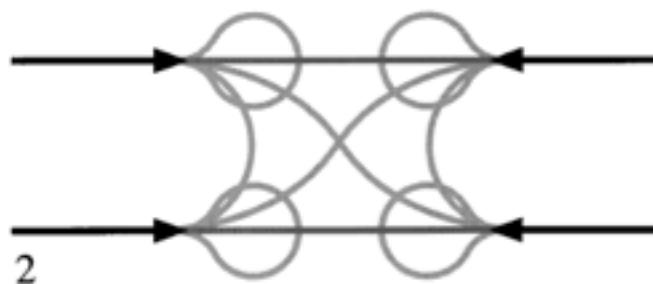
Hum Mutat. 2002 Apr;19(4):410-5.

Padlock Probes
Circularizing Oligonucleotides for Localized DNA Detection
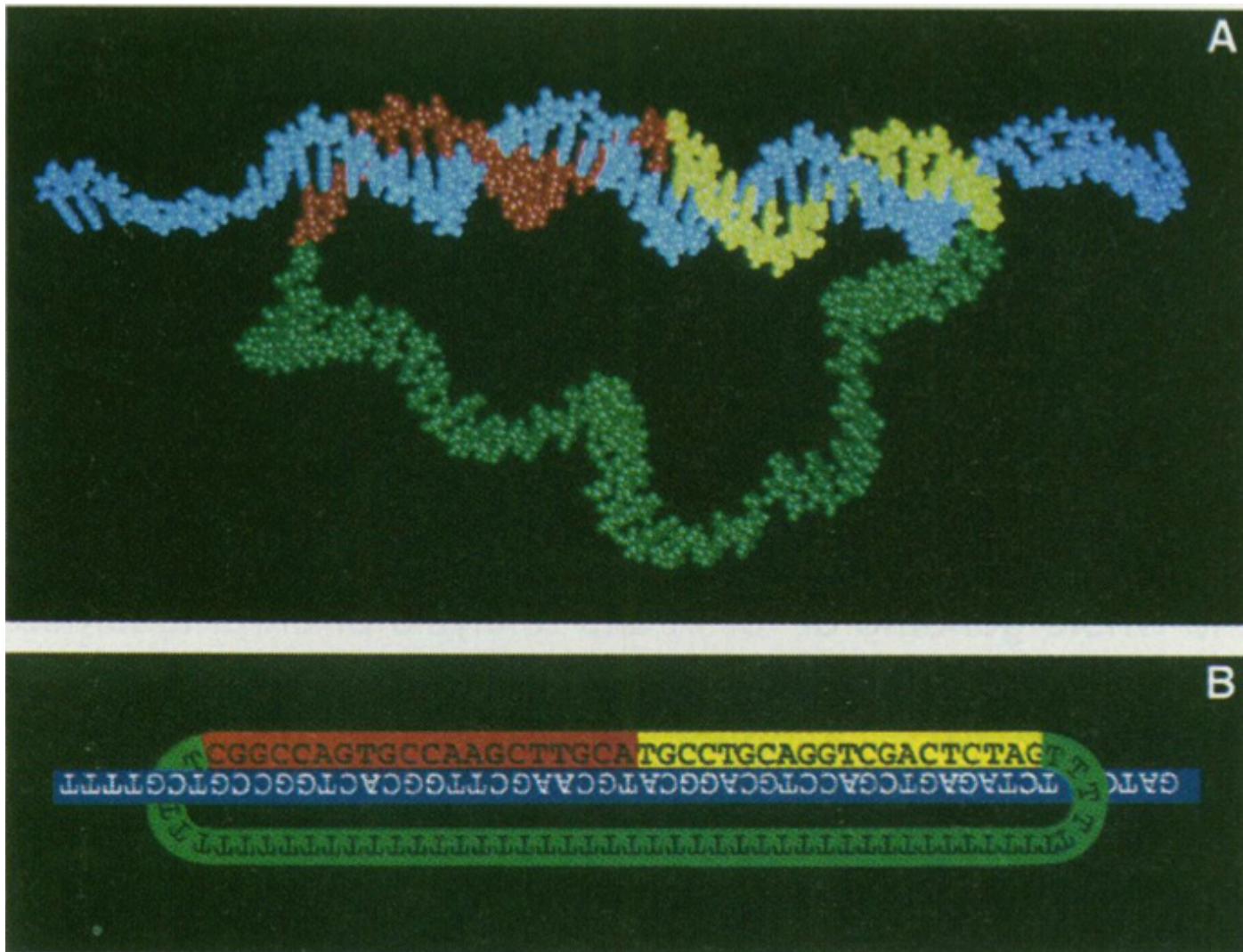
highly parallel molecular bar-coding strategy
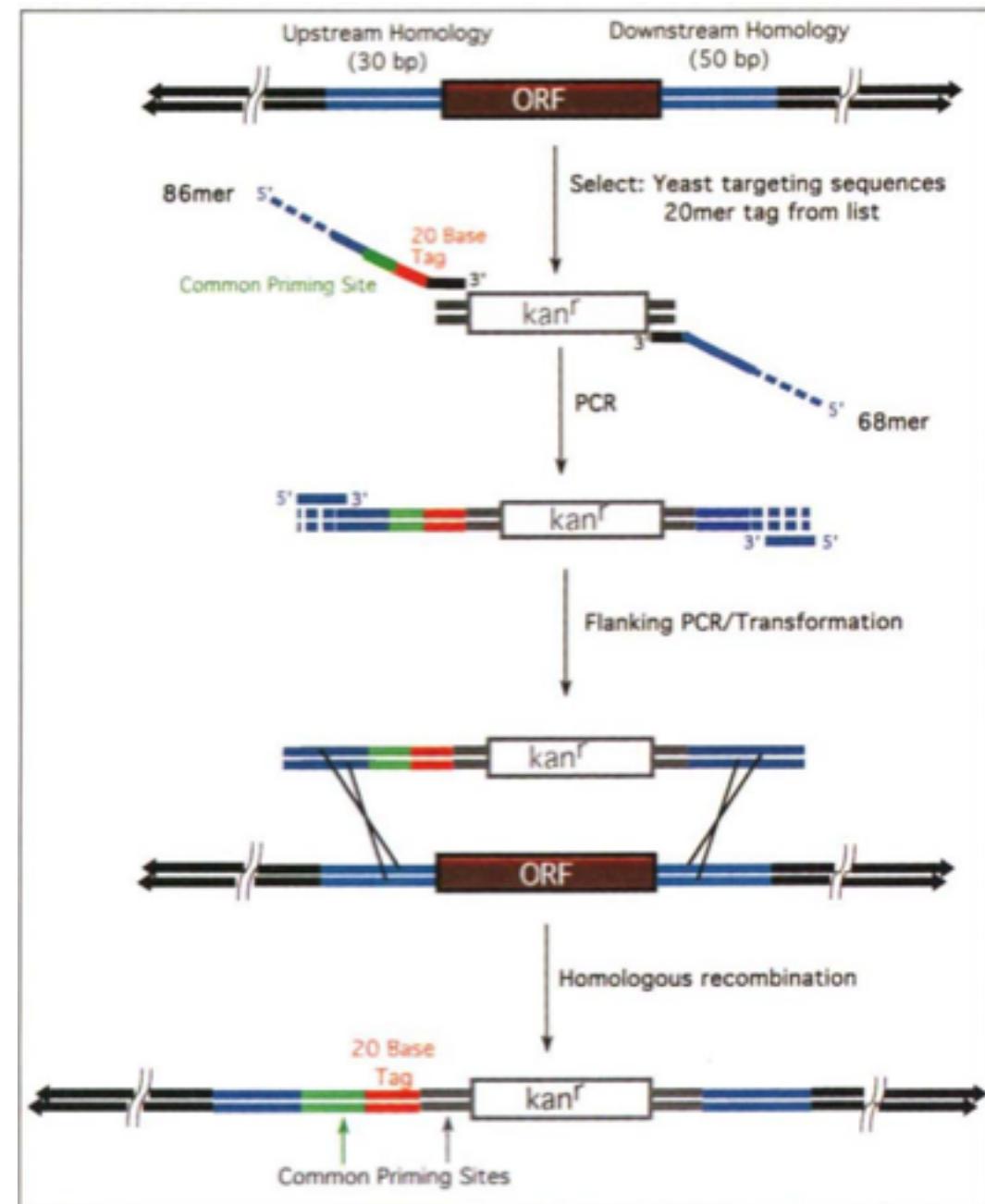
Deletion Pool
(Mixture of 11 different tagged strains)

Genomic DNA

TAG

kan'

3' Labeled primer
(10 fold excess)

PCR (Asymmetric)

3' ——— TAG ——— 5'
56mer Amplicon

Population of 11 different tags

Hybridize to 20mer Array
Wash
Scan

DNATAGS

ADE1	ADE3	ADE5	ARO7	TRP3	TRP5	
	ADE2	ADE4	ARO2	TRP2	TRP4	*
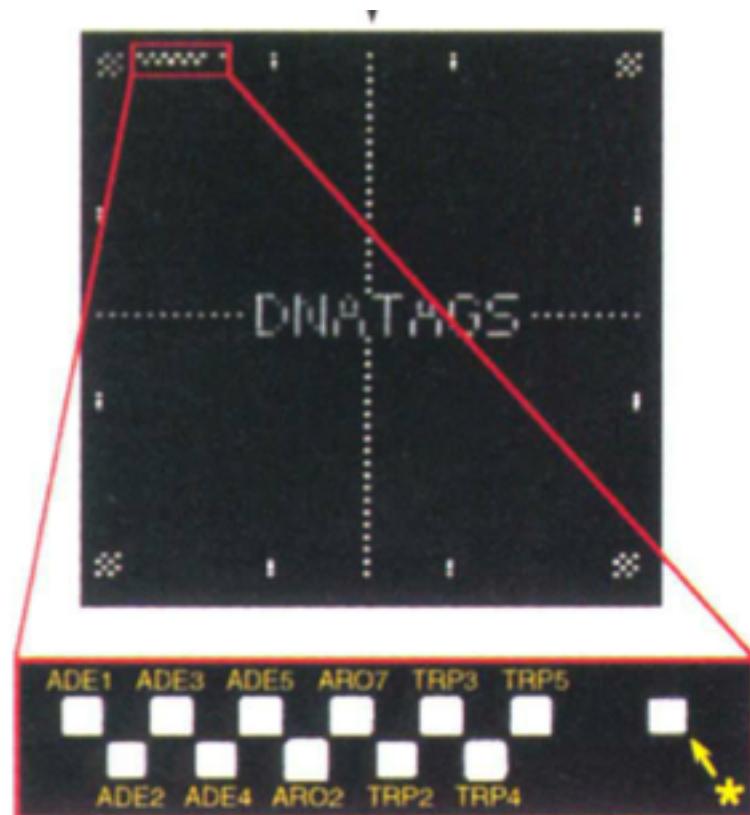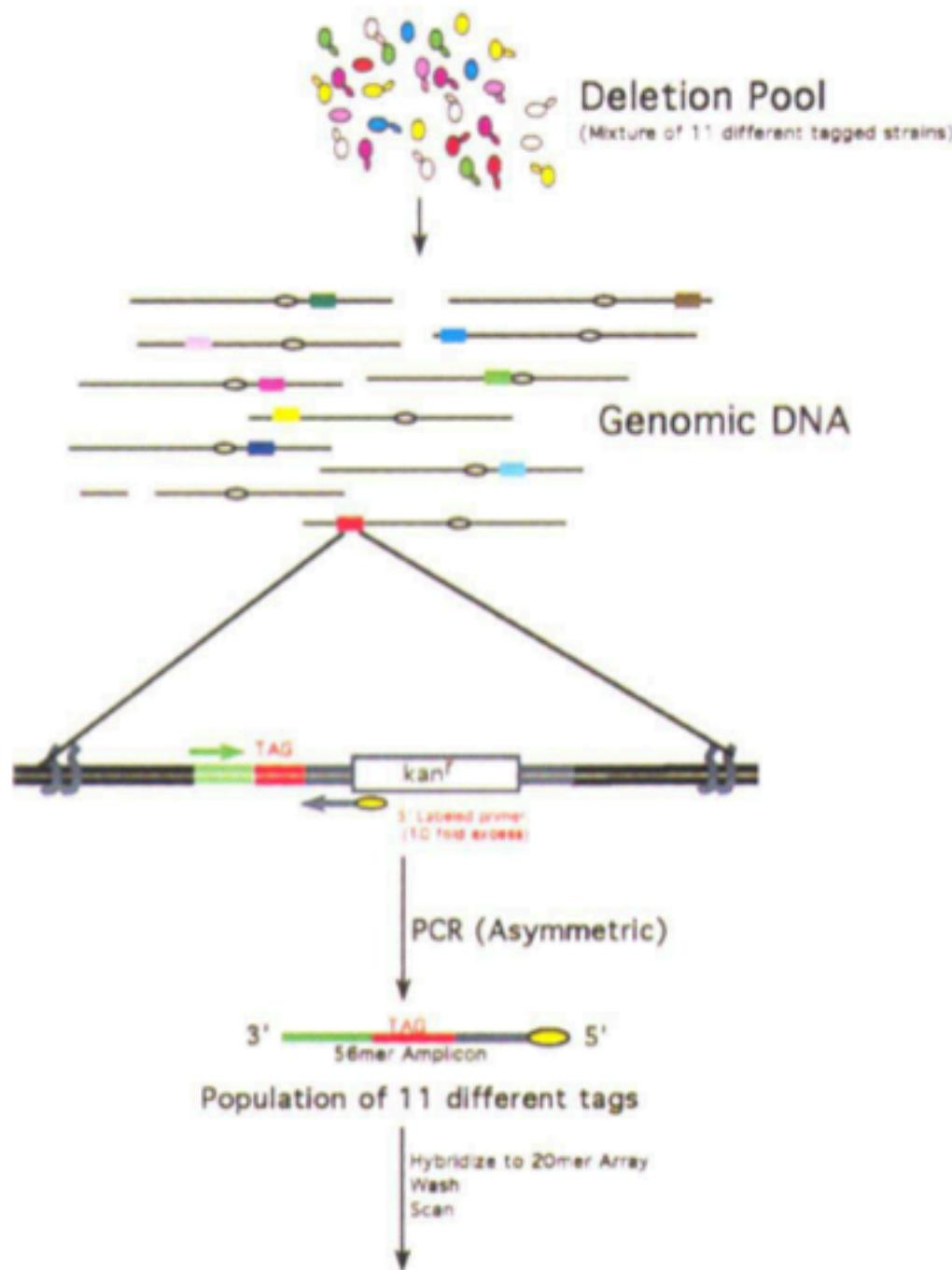
molecular inversion probes-MIP for SNP

MIP Genotyping Process Overview
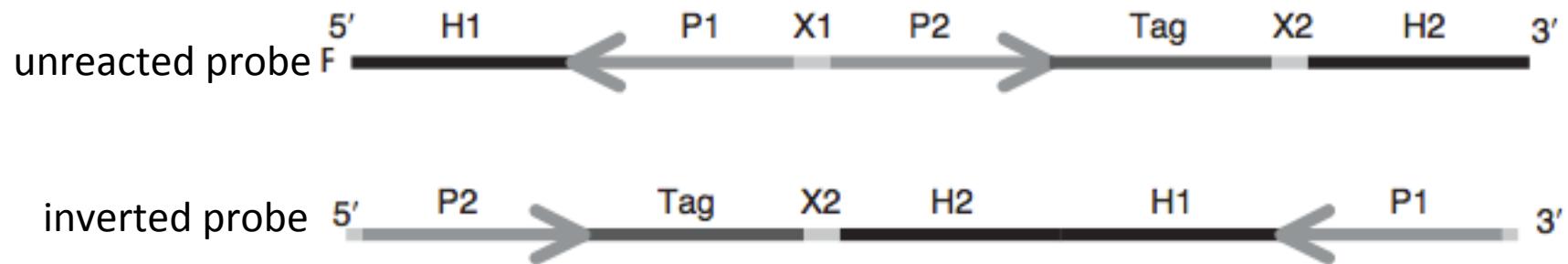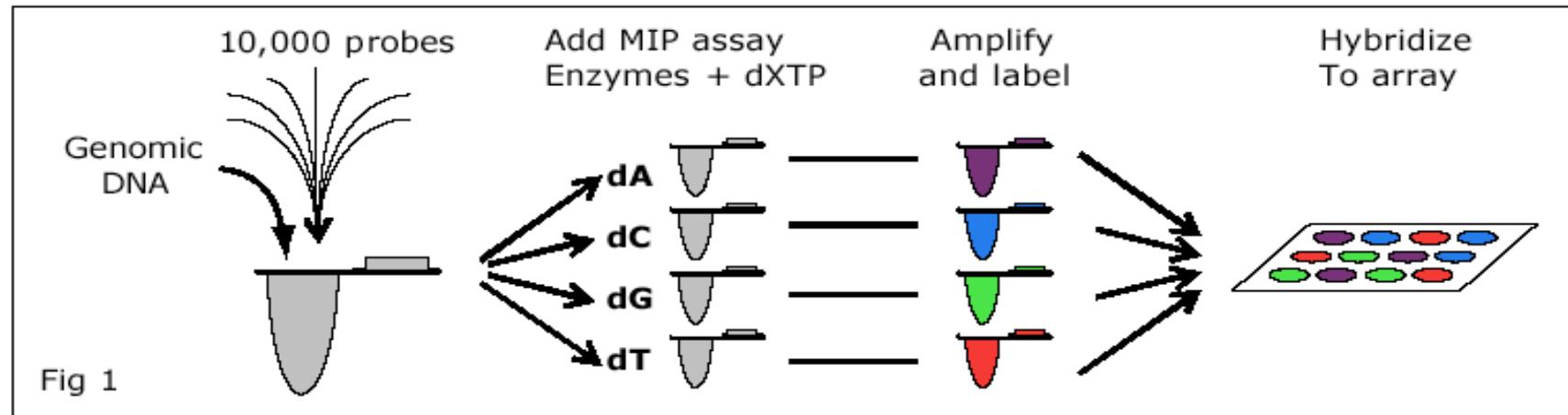
Fig 1

molecular inversion probes-MIP for SNP

1. Anneal

Genomic DNA

C G G A G A T G G C C C A
G C C T C T   C C G G G T
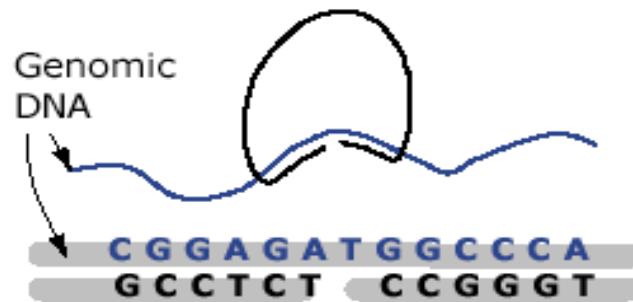
Anneal A mixture of Genomic DNA, up to 10,000 probes, thermostable ligase and polymerase is heat denatured and brought to annealing temperature. Two sequences located at each termini of the probe hybridize to their respective complementary sites on the genome thus forming a circular conformation with a single nucleotide gap between the termini of the probe.

2. Gap Fill - Polymerization

Probe

C G G A G A T G G C C C A
G C C T C T A C C G G G T
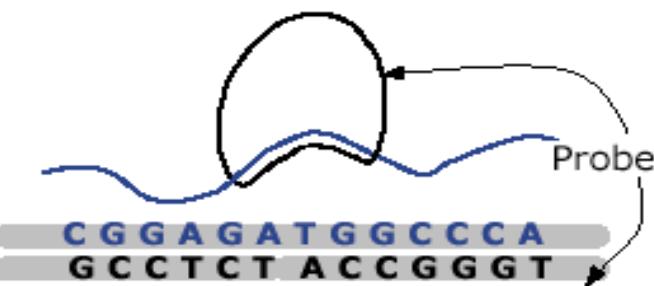
Gap Fill polymerization Unlabeled dATP, dCTP dGTP or dTTP is added to each of the 4 reactions respectively. In reactions where the added nucleotide is complementary to the base being studied, DNA polymerase adds the nucleotide

3. Gap Fill - Ligation

C G G A G A T G G C C C A
G C C T C T A C C G G G T

Gap Fill ligation DNA ligase closes the gap to form a covalently closed circular molecule that encircles the genomic strand to which it is hybridized.

molecular inversion probes-MIP for SNP

4. Exonuclease selection

Exonuclease selection Exonucleases are then added to digest linear probes in reactions where the added nucleotide was not complementary to the gap and excess linear probe in reactions where circular molecules were formed. The reactions are then heated to inactivate the exonucleases.

5. Probe release

Probe release The probes are then cleaved to release them from the genomic DNA

6. Amplification

Amplification The probes are amplified using common primers for all probes

Fig 3

The MIP before, during, and after inversion
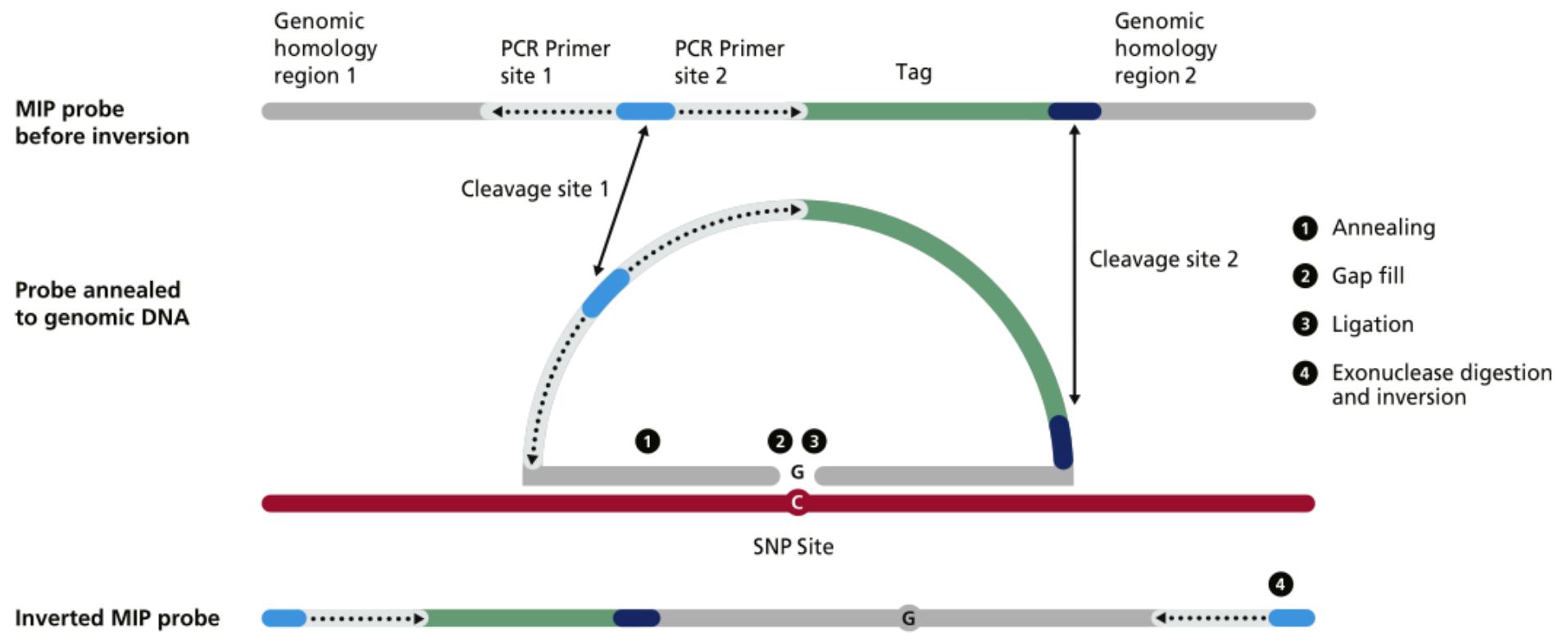
Advantage of MIP

- **precise measurements, large assay dynamic range**.(high specificity & minimum 'cross talk' between loci or alleles)

- **use lower amounts of input genomic DNA**.

- **its use in degraded s**amples, such as fomaldehyde fixed paraffin embedded-FFPE samples (because MIP requires only 40 base-pairs of intact genomic DNA)

evaluating multiplex targeting methods

- key performance parameters to consider include multiplexity, specificity and uniformity.

- Multiplexity refers to the number of independent capture reactions performed simultaneously in a single reaction. Specificity is measured as the fraction of captured nucleic acids that derive from targeted regions. Uniformity is defined as the relative abundances of targeted sequences after selective capture.

- Cost

Modification of the MIP strategy
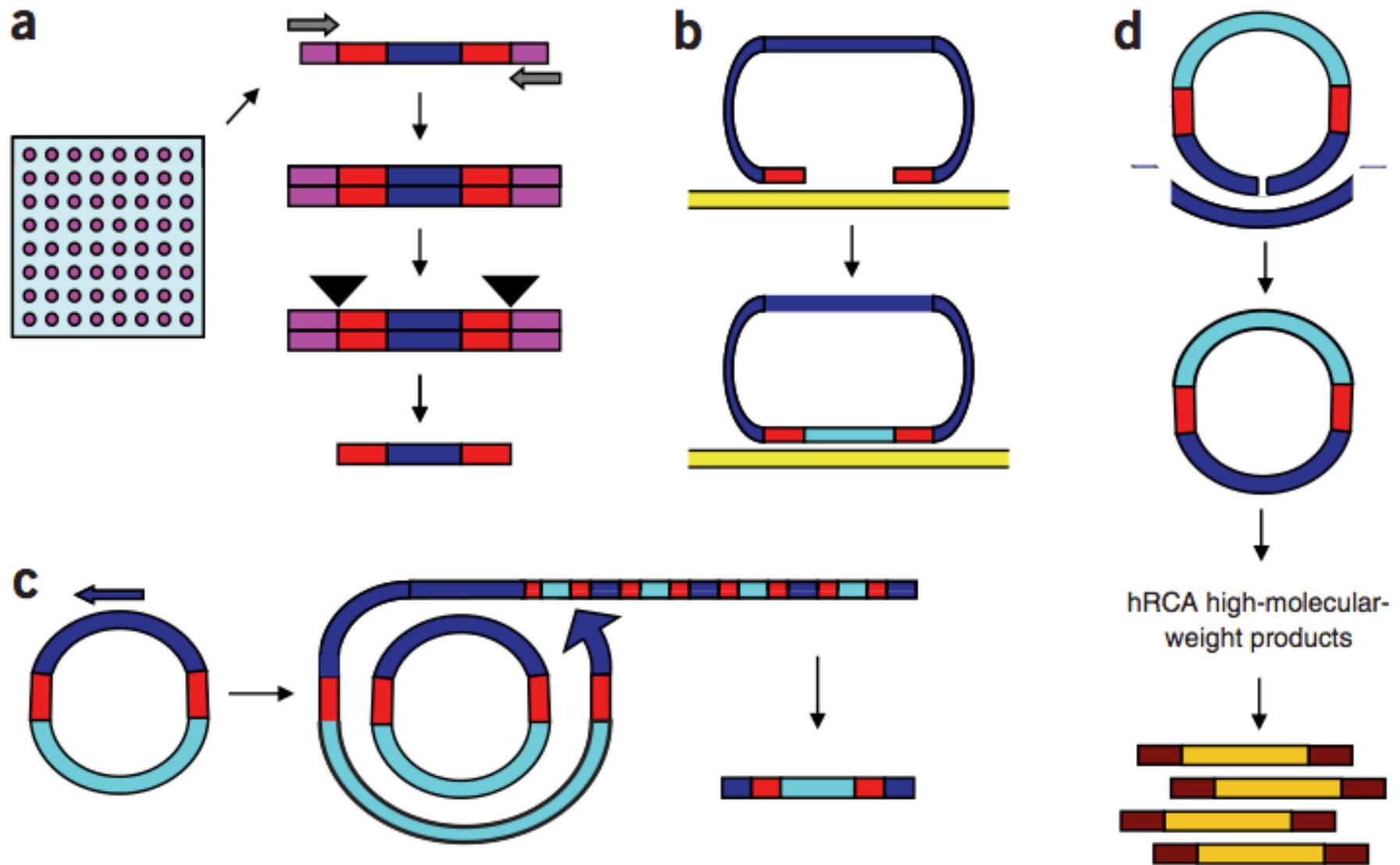
SMART (Spacer Multiplex Amplification Reaction)

long padlock
probes (LPPs):
~500bp gap.
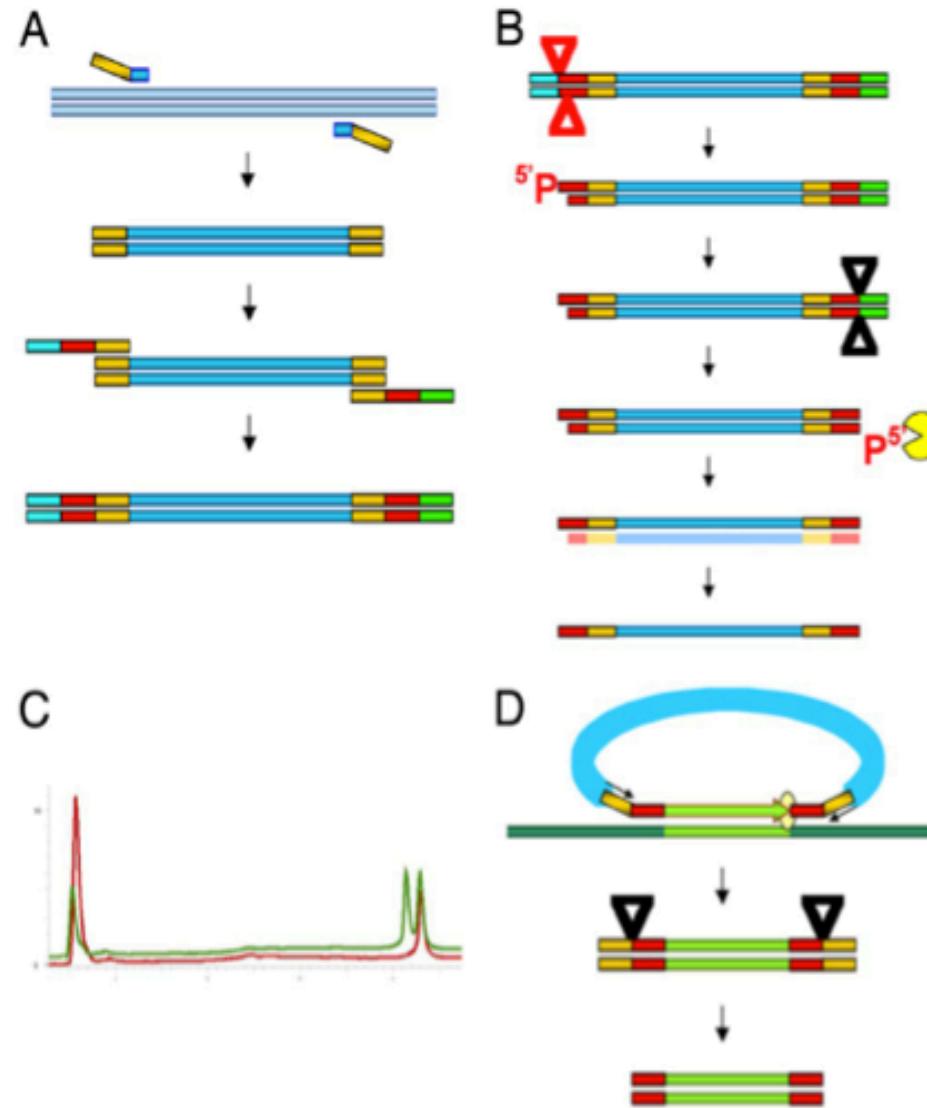
MIP application-1 in resequencing
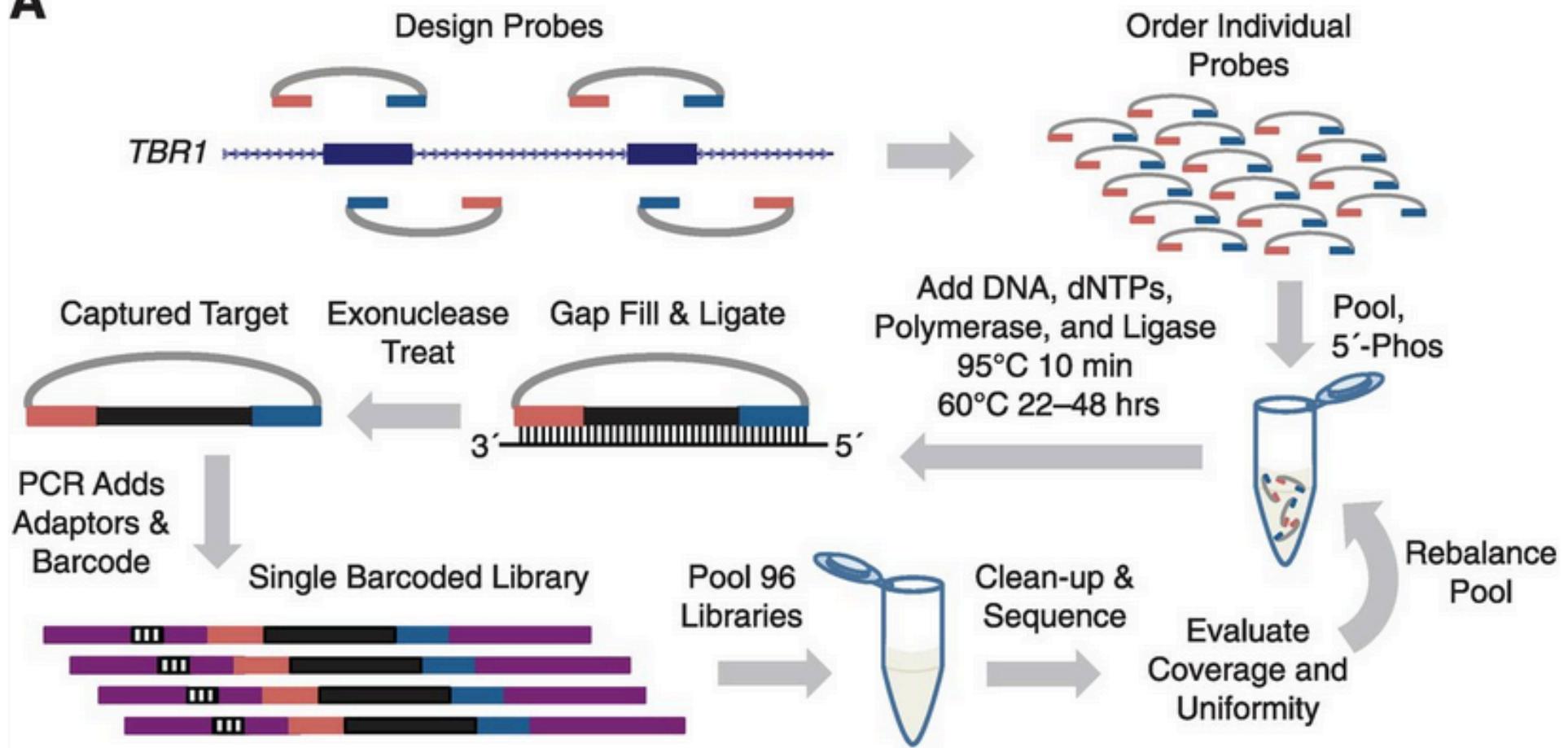
Material & Methods

- 2494 ASD probands from Simon Simplex Collection (2446 successfully capture with both probe sets.

- MIP targeted 44 genes (from 192 candidates): disruptive mutations, associations with syndromic autism, overlap with known or suspected neurodevelopmental CNV, risk loci, structural similarities, and/or neuronal expression.

- 23 of the 44 genes intersect a 49-member beta-catenin–chromatin-remodeling protein-protein interaction (PPI) network or an expanded 74-member network.

- Illumina Hiseq 2000 and Miseq (for confirmations).

49-member PPI network

74-member PPI network

Science. 2012 Dec 21;338(6114):1619-22.

Filtering & Validation

- filtering against variants observed in other cohorts, non-ASD exomes and MIP-based resequencing of 762 healthy, non-ASD individuals.

- Remaining candidates were further tested by MIP-based resequencing of the proband's parents and, if potentially de novo, confirmed by Sanger sequencing of the parent-child trio.

Results - four genes with multiple de novo mutation events
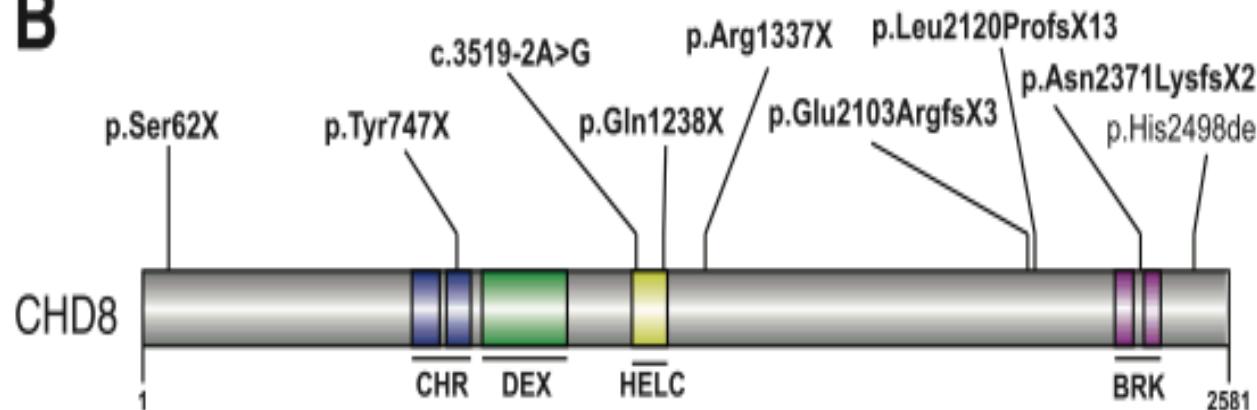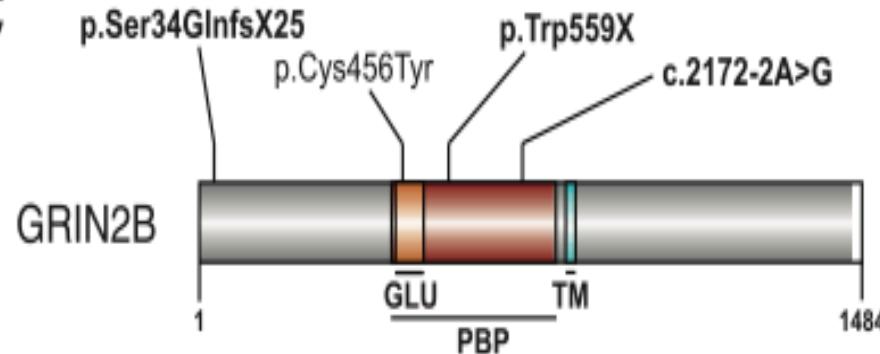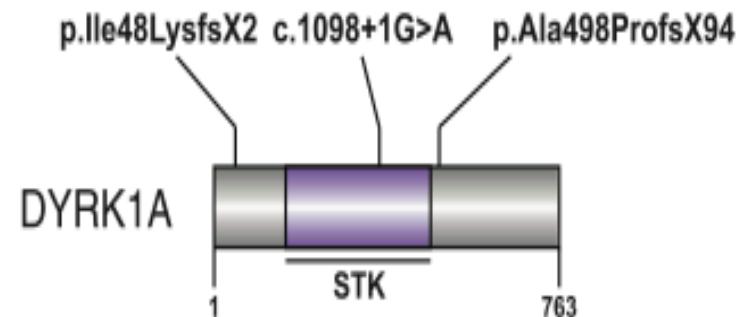
Severe de novo events (bold variants), i.e., coding indels, nonsense mutations, and splice-site disruptions (17/27 or 0.63), is four times the expected proportion for random de novo mutations.

Science. 2012 Dec 21;338(6114):1619-22.

Results - Six genes with recurrent de novo mutations

| Proband	Sex	Gene	Mut	Assay	HGVS	NVIQ
12714.p1	M	CHD8*	Ns	MIP	p.Ser62X	78
13986.p1	M	CHD8*	Fs	MIP	p.Tyr747X	38
11654.p1	F	CHD8*	Sp	MIP‡ (4)	c.3519-2A>G	41
13844.p1	M	CHD8*	Ns	EX	p.Gln1238X	34
14016.p1	M	CHD8*	Ns	MIP	p.Arg1337X	92
12991.p1	M	CHD8*	Fs	MIP	p.Glu2103ArgfsX3	67
12752.p1	F	CHD8*	Fs	EX	p.Leu2120ProfsX13	93
14233.p1	M	CHD8*	Fs	MIP	p.Asn2371LysfsX2	19
14406.p1	M	CHD8*	Aa	MIP	p.His2498del	98
12099.p1	M	DYRK1A*	Fs	MIP‡ (4)	p.Ile48LysfsX2	55
13890.p1	F	DYRK1A*	Sp	EX	c.1098+1G>A	42
13552.p1	M	DYRK1A*	Fs	MIP§ (6)	p.Ala498ProfsX94	66
11691.p1	M	GRIN2B†	Fs	MIP‡‖ (3)	p.Ser34GlnfsX25	62
13932.p1	M	GRIN2B†	Ms	MIP	p.Cys456Tyr	55
12547.p1	M	GRIN2B†	Ns	MIP‖	p.Trp559X	65
12681.p1	F	GRIN2B†	Sp	EX	c.2172-2A>G	65
14433.p1	M	PTEN	Ms	MIP	p.Thr131Ile	50
14611.p1	M	PTEN	Fs	MIP	p.Cys136MetfsX44	33
11390.p1	F	PTEN	Ms	EX	p.Thr167Asn	77
12335.p1	F	TBL1XR1*	Ms	EX	p.Leu282Pro	47
14612.p1	M	TBL1XR1*	Fs	MIP	p.Ile397SerfsX19	41
11480.p1	M	TBR1†	Fs	EX	p.Ala136ProfsX80	41
13814.p1	M	TBR1†	Ms	MIP	p.Lys228Glu	78
13796.p1	F	TBR1†	Fs	MIP‡ (4)	p.Ser351X	63

*Part of 49-member connected component reported in (3).   †Part of expanded 74-member connected component.   ‡,§Proband was exome sequenced by cited study and variant was ‡not reported or §reported.   ‖Variant reported in MIP screen from (3).

Results-mutation burden

A

"Probabilities shown are for observing x or more events, of which at least y belong to the severe class."

$-\log_{10}(p\text{-value})$

Genes: Total, CHD8*, GRIN2B*, PTEN, DYRK1A*, TBR1*, TBL1XR1*, CTNNB1*, ADNP*, ARID1B*, UBE3C*, SGSM3, ADCY5*, SCN1A*, SETD2, SBF1, LAMC3
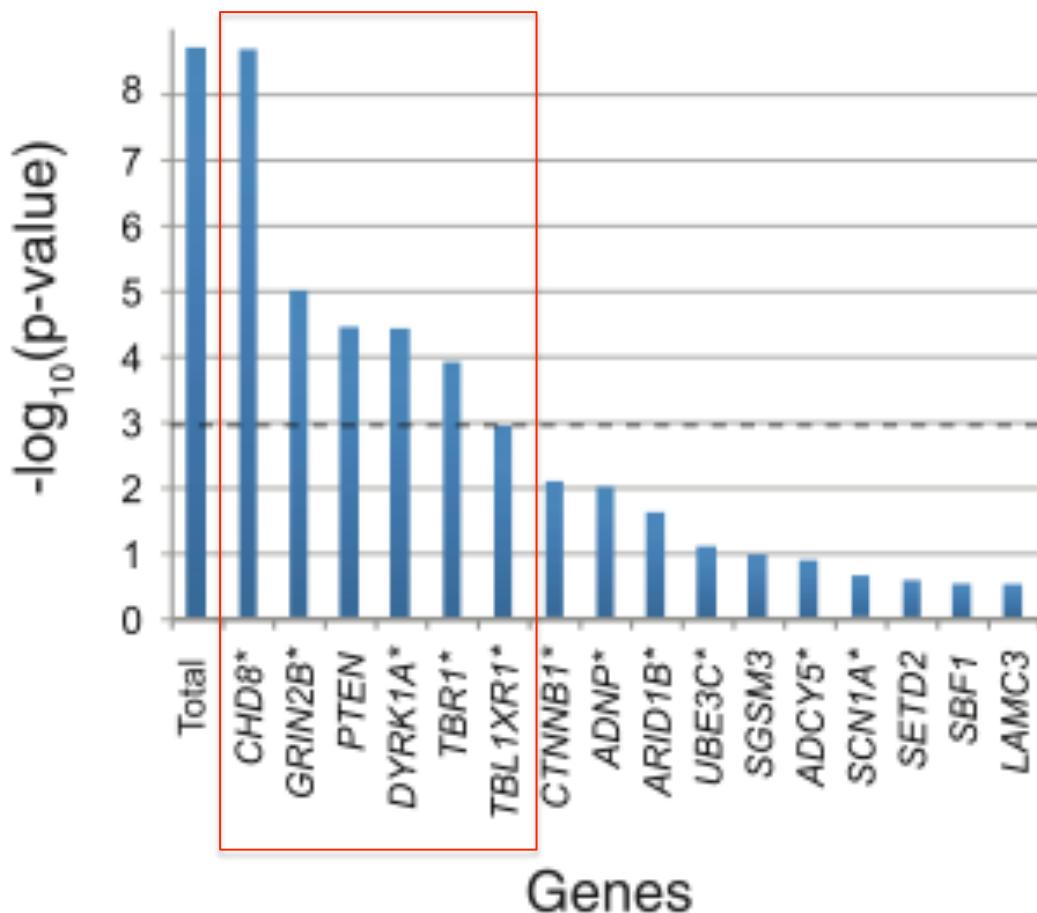
- Higher rate of de novo mutation than expected—in the overall set of 44 genes, driven by the severe class (16/17 intersect 74 PPI network).

- *CHD8, GRIN2B, DYRK1A, PTEN, TBR1, and TBL1XR1,* five fall within the beta-catenin–chromatin-remodeling network.

- ~1% of ASD probands harbor a de novo mutation in one of these six genes. (*CHD8*-0.35%)

Science. 2012 Dec 21;338(6114):1619-22.

Results

Table S5. Comparison of ASD1 MIP variant calls to exome variant calls for 48 samples (16 trios).

	Total	in dbSNP	% in dbSNP	Not in dbSNP	% Novel
Called by both	725	699	96.4	26*	3.6
Called only by MIP	72	63	87.5	9	12.5
Called only by exome	7	7	100	0	0

*All *de novo* events called by exome sequencing were also called by MIP-based resequencing.

Results – head circumference

microcephaly in our index DYRK1A mutation case, macrocephaly in both probands with CHD8 mutations.
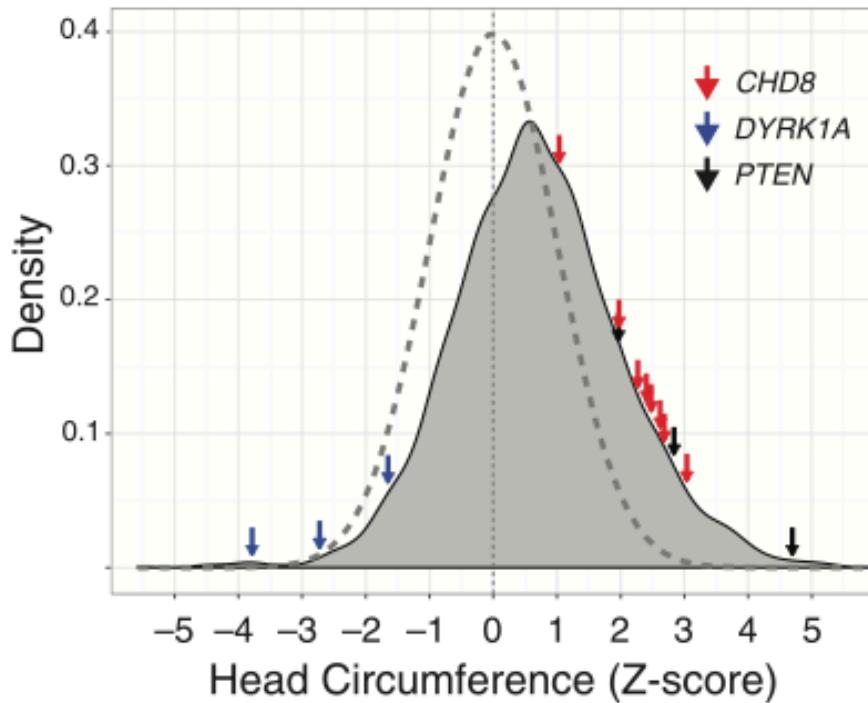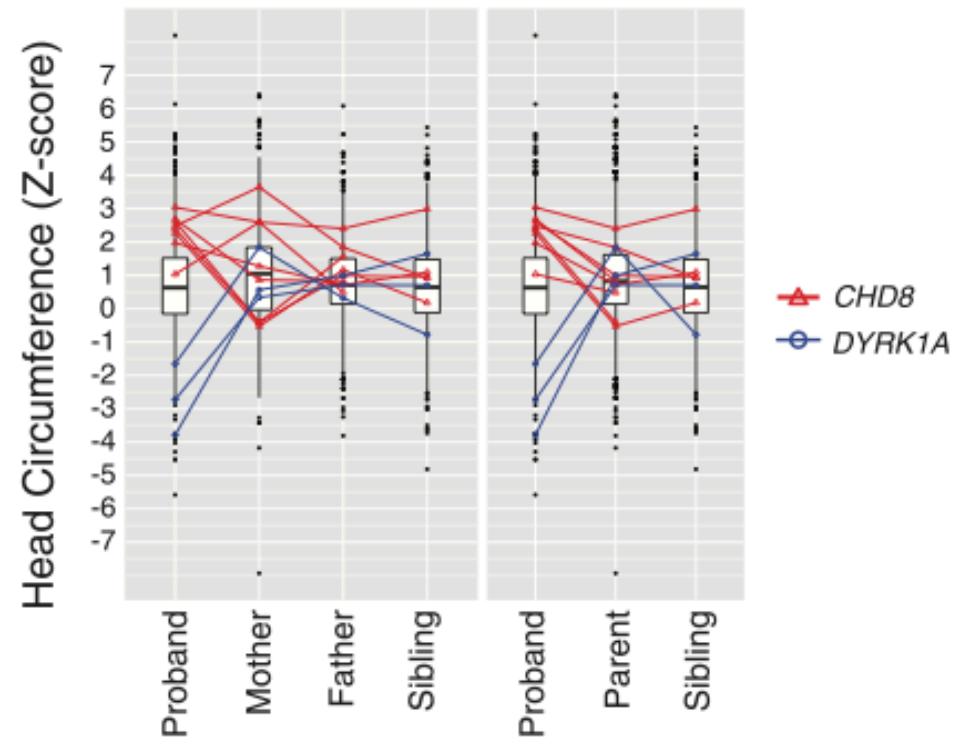
De novo CHD8 mutations are present in ~2% of macrocephalic (HC > 2.0) SSC probands, a useful phenotype for patient subclassification.

Science. 2012 Dec 21;338(6114):1619-22.

Conclusion

- Our data support an important role for de novo mutations in six genes in ~1% of sporadic ASDs.

- "Whereas our data implicate specific loci in ASDs, they are insufficient to evaluate whether the observed de novo mutations are sufficient to cause ASDs."

MIP application-2 in resequencing- smMIP
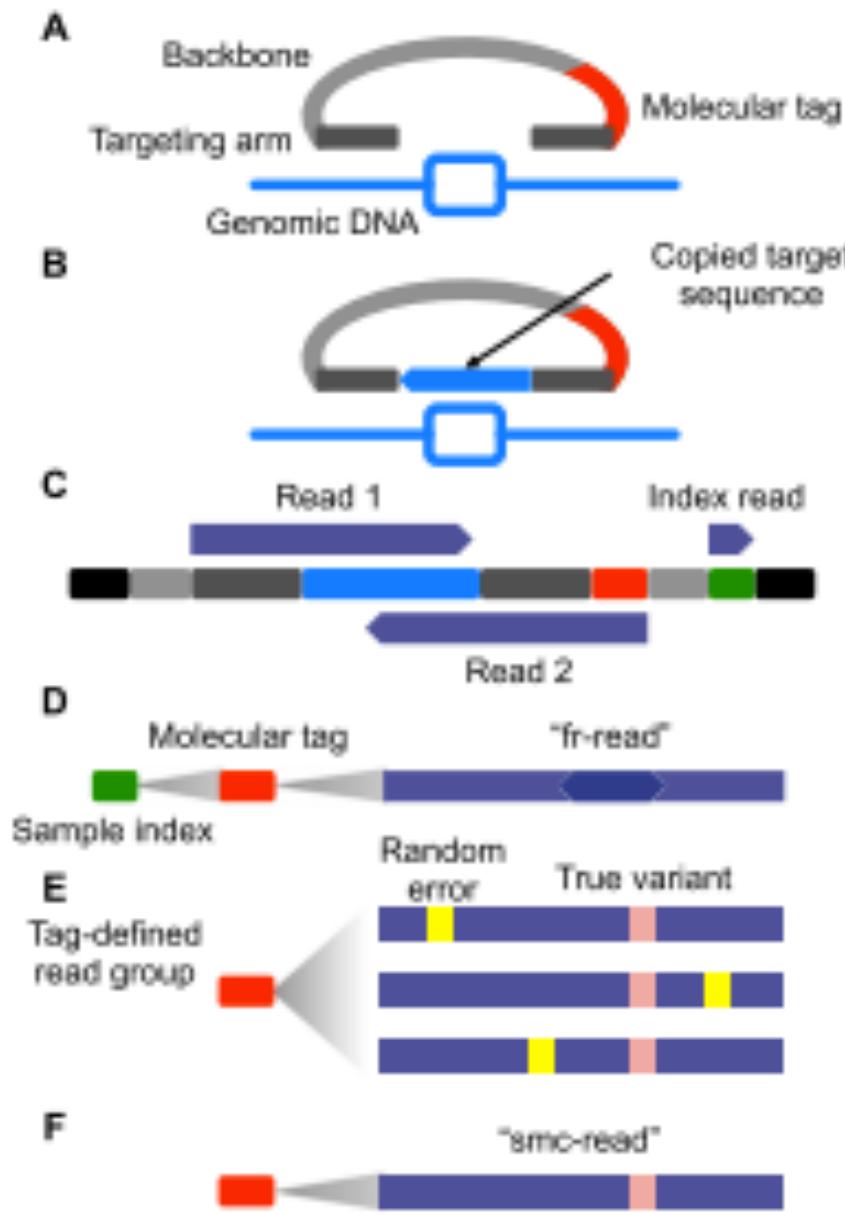
''sample index'': resolving capture products from distinct source DNAs;

''molecular tag'': resolving reads derived from distinct genomic equivalents within individual source DNAs.

''fr-reads'': Before alignment, overlapping regions of read-pairs are reconciled to produce ;152-nt forward- reverse reads;

''smc-reads'': After alignment, groups of fr-reads form the basis for highly accurate single molecule consensus reads.

Genome Res. 2013 Mar 11.

Material & Methods

- Design 1312 smMIP oligonucleotides.

- 53 specimens: 45 clinical cancer specimens+8 HapMap DNA mixtures (two HapMap cell lines).

- Tiling ~125 Kb of genomic sequence, including 99% of coding bases of 33 cancer-related genes.

- Massively parallel sequencing- Illumina Hiseq 2000 and Miseq, and analyzed using a custom pipeline.

Results-Coverage Distributions

A

Genome Res. 2013 Mar 11.

- Mean smc-read coverage: HapMap samples: 3538x; Clinical specimens: 1051x.

- >100× smc-read coverage: 78% of the clinical samples and all HapMap cell line samples, at least 85% of all targeted coding bases.

- >1,000× smc-read coverage: all of the HapMap samples, 60% of all targeted coding bases; 42% of the clinical samples had at least 50% of targeted coding bases.

- >10× smc-read coverage: all HapMap samples and 80% of the clinical samples had at least 97% of targeted coding bases.

Results – Subclonal variant detection
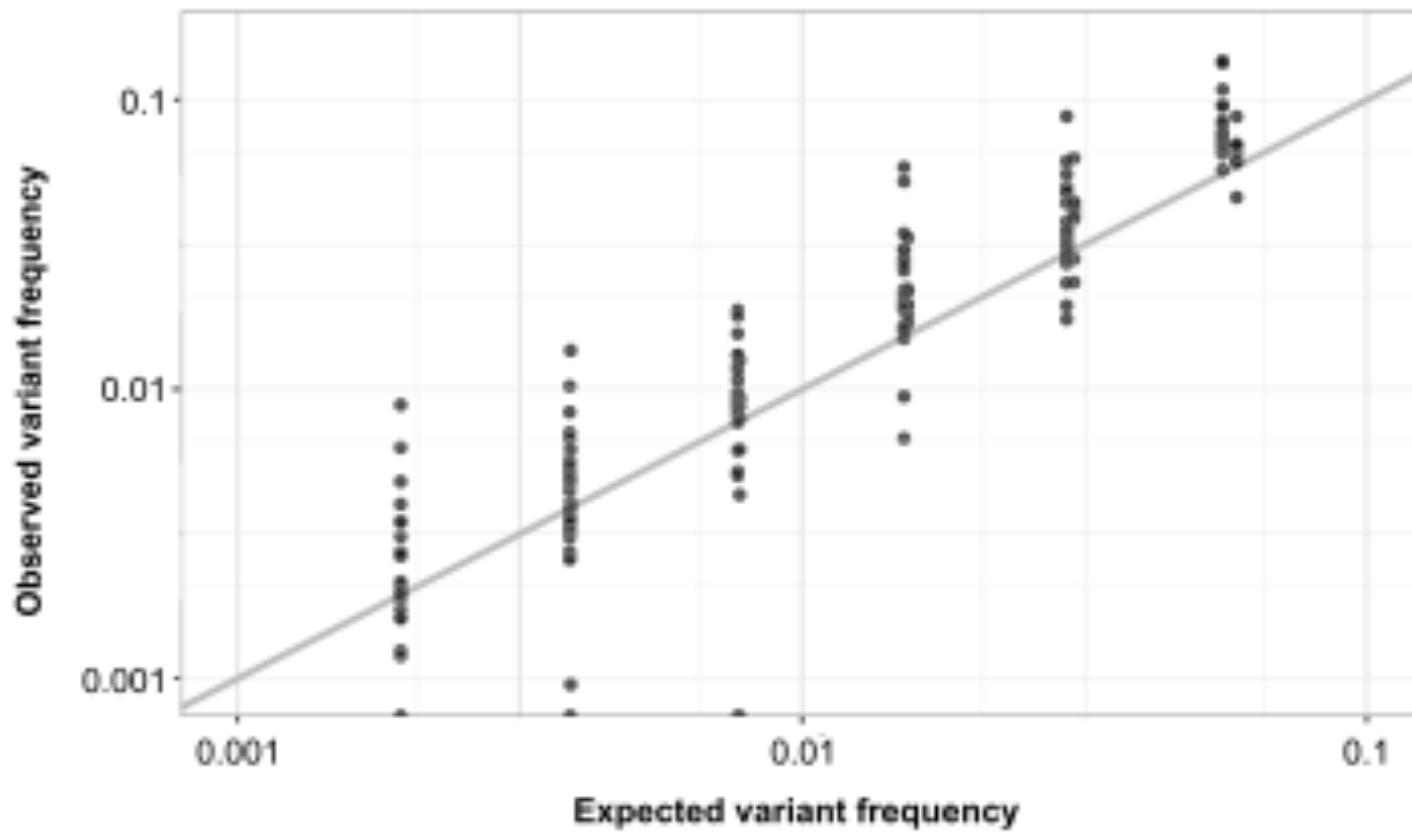
Observed vs expected variant frequency
in smc-read base-calls from mixtures of HapMap
samples, for positions with >100x smc-read coverage.

Genome Res. 2013 Mar 11.

Results – Substitution error rates

Table 2. Substitution error rates

		fr-reads		smc-reads		
		Calls	Sub. rate	Calls	Sub. rate	Fold-reduction in sub. rate
HapMap cell lines	All	1.0×10^{10}	1.1×10^{-4}	4.6×10^{9}	8.4×10^{-6}	12.8
	No G>A, C>A	8.8×10^{9}	1.0×10^{-4}	3.9×10^{9}	3.5×10^{-6}	28.8
Clinical (fresh)	All	2.3×10^{9}	1.1×10^{-4}	6.6×10^{8}	9.5×10^{-6}	11.5
	No G>A, C>A	2.0×10^{9}	1.0×10^{-4}	5.6×10^{8}	2.9×10^{-6}	34.6
Clinical (FFPE)	All	2.0×10^{10}	1.3×10^{-4}	7.1×10^{9}	2.9×10^{-5}	4.5
	No G>A, C>A	1.7×10^{10}	1.0×10^{-4}	6.0×10^{9}	5.3×10^{-6}	19.8

Gap-fill

```
5'------?--->  3'
3'------X------5'
```
Genomic DNA

5' dinuc
```
5'------??--->  3'
3'------X------5'
```

3' dinuc
```
5'------??-->  3'
3'------X------5'
```

PS: Excluding the G>A and C>A substitutions that are likely caused at least in part by patterns of DNA damage (deamination of C and 5-methyl-C, and oxidative damage to G resulting in 8-oxo-G).

Genome Res. 2013 Mar 11.

Results – Substitution error rates

the occurrence of pro-mutagenic chemical processes in individual progenitor molecules:
- C>A: the oxidatively damaged base 8-oxo-G
- G>A: spontaneous deamination of C and 5-methyl- C

FPE treatment may in-crease 8-oxo-G formation rates.

Genome Res. 2013 Mar 11.

Results – Substitution error rates
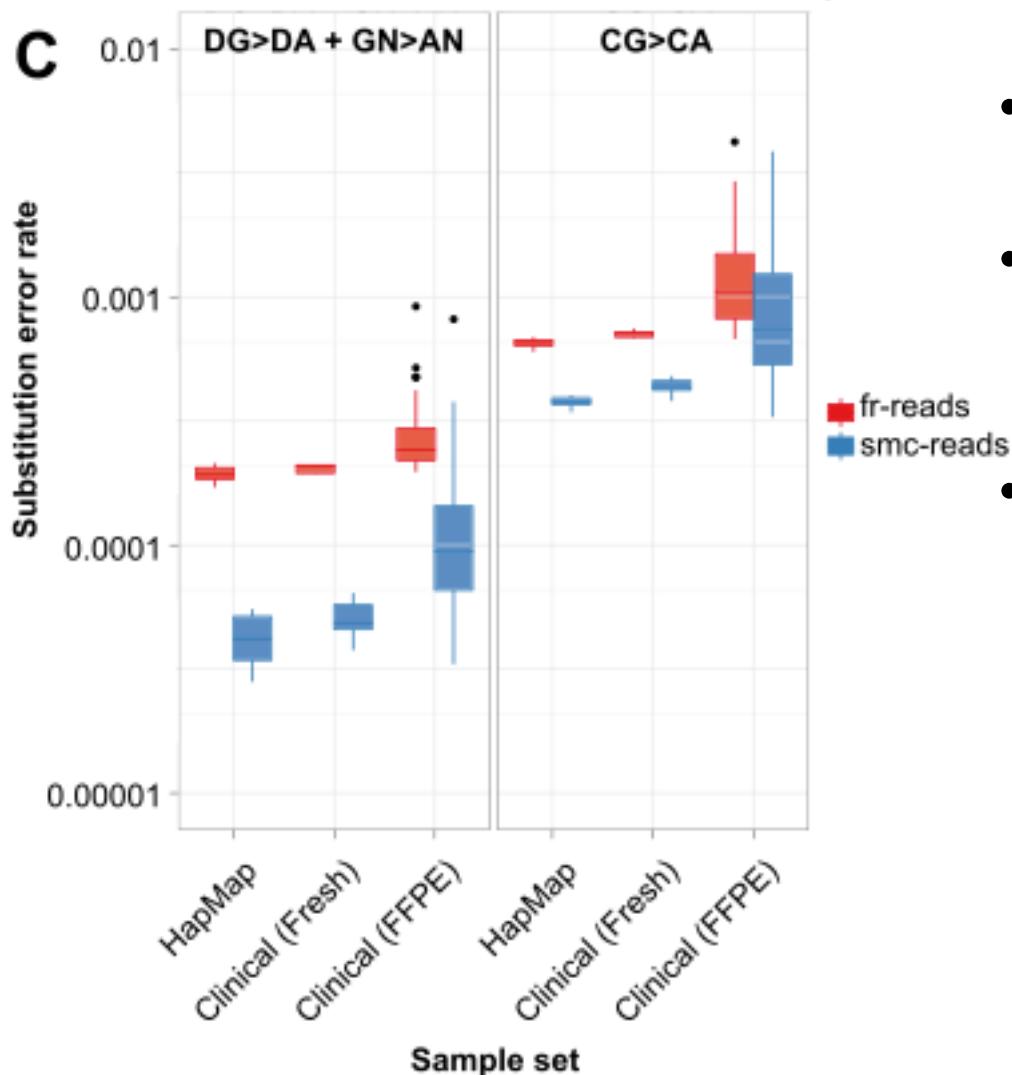
- CG is an infrequent dinucleotide
- the substitution rate of G>A in the absence of CG>CA remains elevated
- deamination of 5mC is not the only factor causing high rates of G>A substitution, and that deamination of C to U may also be playing a substantial role.

Genome Res. 2013 Mar 11.

Results – Substitution error rates

Table 2. Substitution error rates

		fr-reads		smc-reads		
		Calls	Sub. rate	Calls	Sub. rate	Fold-reduction in sub. rate
HapMap cell lines	All	1.0×10^{10}	1.1×10^{-4}	4.6×10^{9}	8.4×10^{-6}	12.8
	No G>A, C>A	8.8×10^{9}	1.0×10^{-4}	3.9×10^{9}	3.5×10^{-6}	28.8
Clinical (fresh)	All	2.3×10^{9}	1.1×10^{-4}	6.6×10^{8}	9.5×10^{-6}	11.5
	No G>A, C>A	2.0×10^{9}	1.0×10^{-4}	5.6×10^{8}	2.9×10^{-6}	34.6
Clinical (FFPE)	All	2.0×10^{10}	1.3×10^{-4}	7.1×10^{9}	2.9×10^{-5}	4.5
	No G>A, C>A	1.7×10^{10}	1.0×10^{-4}	6.0×10^{9}	5.3×10^{-6}	19.8

PS: Excluding the G>A and C>A substitutions that are likely caused at least in part by patterns of DNA damage (deamination of C and 5-methyl-C, and oxidative damage to G resulting in 8-oxo-G).

Genome Res. 2013 Mar 11.

# Results – Sensitivity and false discovery rate
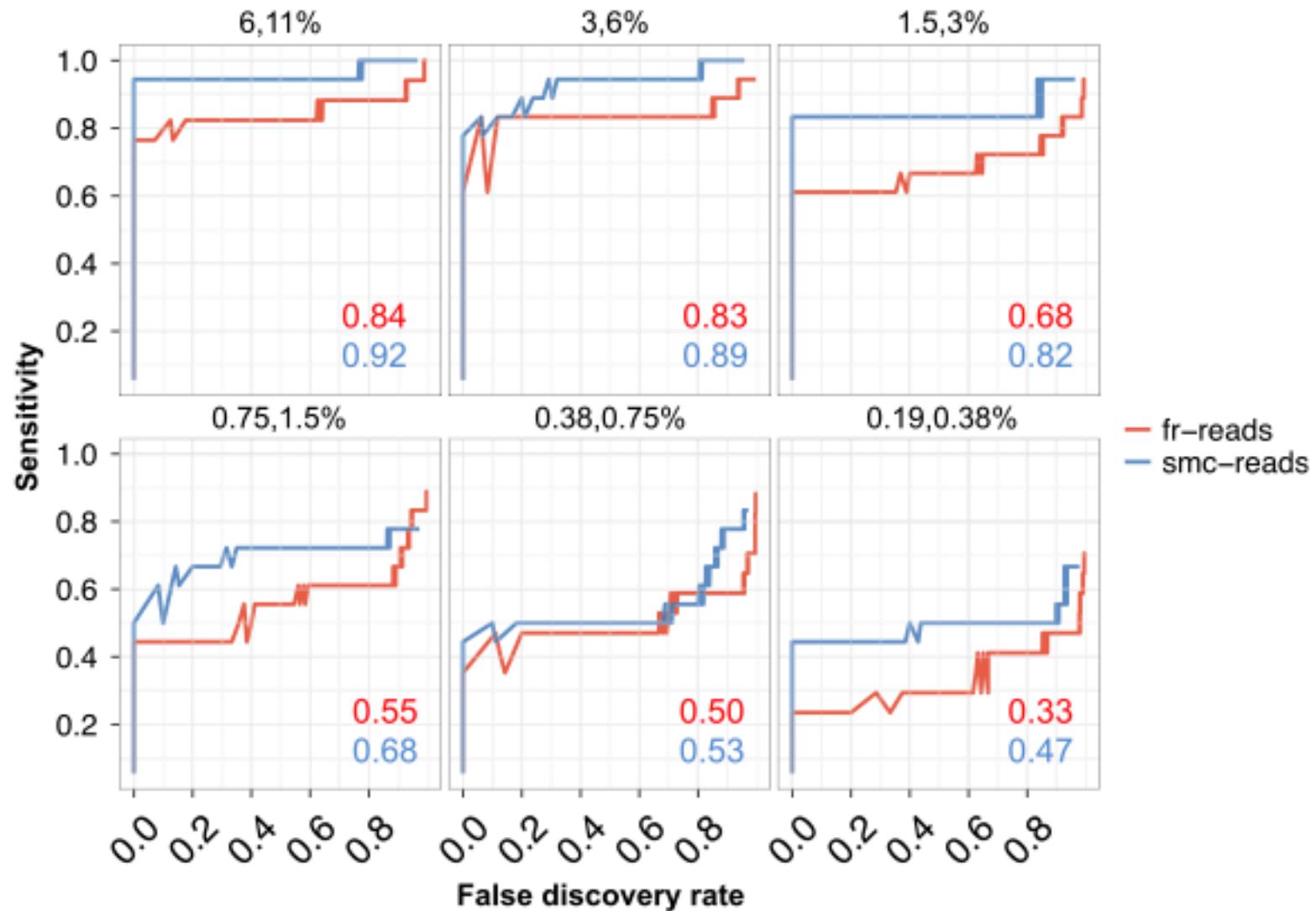
Results – Detection of somatic variation

Table 3. Concordance with single mutation tests

Gene	Mutation	Expected # of events	% detected
BRAF	p.V600E	4	100
BRAF	p.V600K	2	100
EGFR	p.L858R	2	100
EGFR	15-bp deletion (exon 19)	1	100
EGFR	18-bp deletion (exon 19)	1	100
FLT3	67-bp insertion	1	0
FLT3	104-bp insertion	1	0
JAK2	p.V617F	3	100
KIT	6-bp insertion (exon 11)	1	100
KIT	15-bp deletion (exon 11)	1	100
KRAS	p.G12C	2	100[a]
KRAS	p.G12V	1	100[a]
KRAS	p.G12D	2	100
KRAS	p.G13D	2	100
NRAS	p.Q61R	1	100
PDGFRA	p.D842V	2	100
Total	All	27	92.6%

Genome Res. 2013 Mar 11.

Results- Alternate allele frequencies of somatic variants in clinical samples (tumor subclones)

A
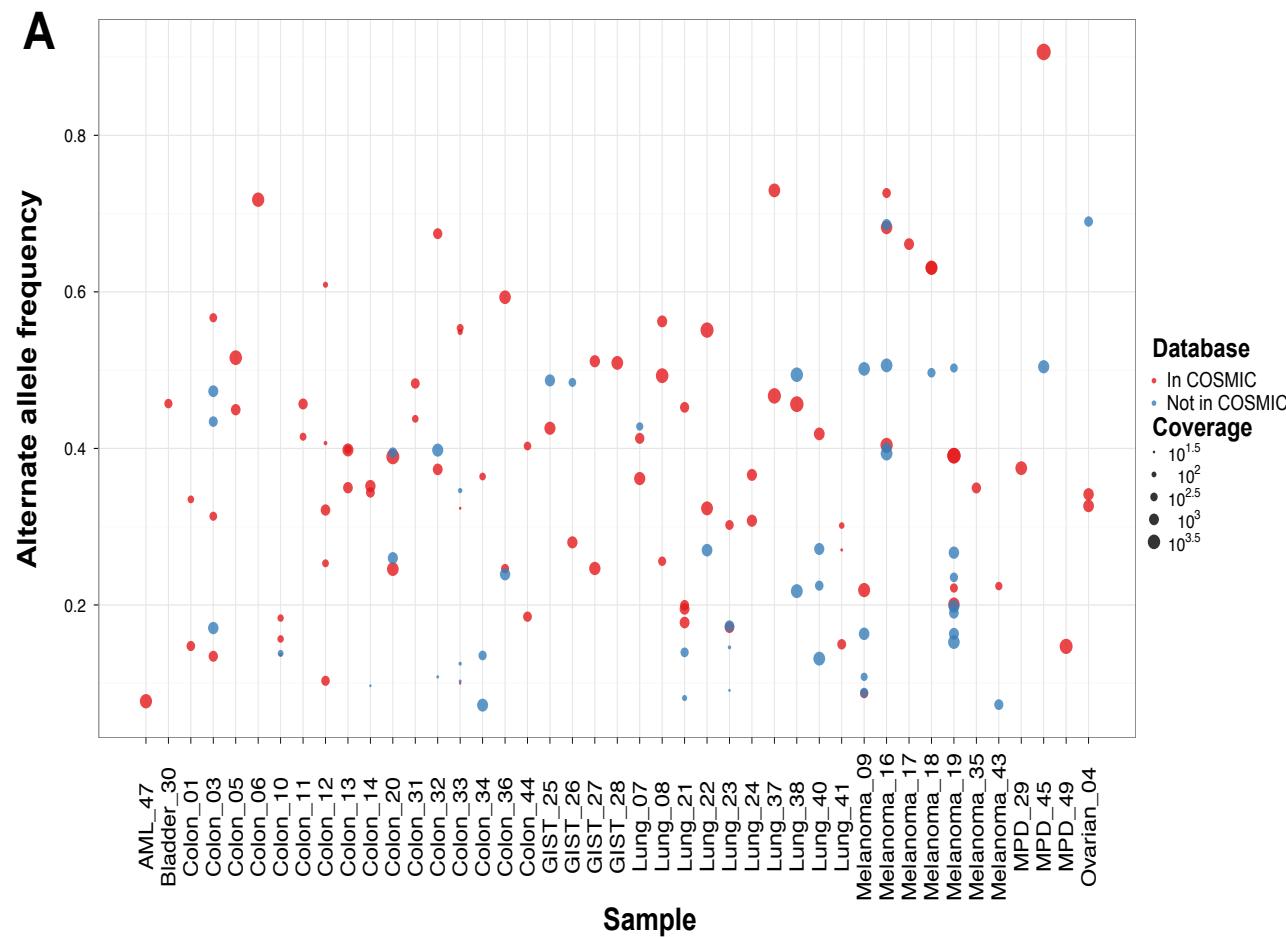
"smMIP assay is highly quantitative for alternate allele frequencies as low as ~0.2%."

Genome Res. 2013 Mar 11.

Results – subclonal somatic variztion at clinically informative sites

Table 4. Low-frequency variation at clinically relevant sites in tumor samples

Sample	Cancer type	Gene	Chr.	Pos.	Ref. allele	Alt. allele	Sub. in gap-fill	Ref. allele counts	Alt. allele counts	Alt. allele fraction	Mutation in protein
43	Melanoma	BRAF	7	140453136	A	T	A>T, T>A	1266	3	0.0024	p.V600E
19	Melanoma	JAK2	9	5073770	G	T	G>T, C>A	1465	19	0.0128	p.V617F
34	Colon	JAK2	9	5073770	G	T	G>T, C>A	1058	15	0.0140	p.V617F
38	Lung	JAK2	9	5073770	G	T	G>T	1229	4	0.0032	p.V617F
41	Lung	JAK2	9	5073770	G	T	G>T, C>A	795	6	0.0075	p.V617F
41	Lung	KRAS	12	25398284	C	A	G>T	464	3	0.0064	p.G12V
12	Colon	NRAS	1	115256529	T	C	A>G	184	9	0.0466	p.Q61R

$P < 10^{-7}$.

"However, based on the FDR analysis performed using the HapMap samples, the P-value cutoff of 10^{-7} is expected to yield a nontrivial FDR of 40% for variants near 0.1% frequency, so it remains possible that one or more of these variants is artifactual."

Genome Res. 2013 Mar 11.

Results - Rapid workflow timetable (miseq)

Step nbr.	Step description	Time (min)	Time (hrs)	Total time (mins)	Total time (hrs)	Start day	End day	Start time	End time
1	Isolate genomic DNA	240	4.00	240	4.00	1	1	9:00 AM	1:00 PM
2	Hybridization, Gap-fill, Ligation*	300	5.00	540	9.00	1	1	1:00 PM	6:00 PM
3	Wait overnight	900	15.00	1440	24.00	1	2	6:00 PM	9:00 AM
4	Exonuclease	65	1.08	1505	25.08	2	2	9:00 AM	10:05 AM
5	PCR	120	2.00	1625	27.08	2	2	10:05 AM	12:05 PM
6	SPRI purification (1.8x)	20	0.33	1645	27.42	2	2	12:05 PM	12:25 PM
7	SPRI purification (0.8x)	20	0.33	1665	27.75	2	2	12:25 PM	12:45 PM
8	Gel	60	1.00	1725	28.75	2	2	12:45 PM	1:45 PM
9	Pooling and quantification	20	0.33	1745	29.08	2	2	1:45 PM	2:05 PM
10	MiSeq	1680	28.00	3425	57.08	2	3	2:05 PM	6:05 PM
11	Analysis**	360	6.00	3785	63.08	3	3	6:05 PM	12:05 AM

Results-Performance of rapid workflow

Sample number	Cancer type	Fraction of targeted bases covered	Fraction of high-coverage sites covered	Concordance with high-coverage data	Clinically informative mutation	Detected clinically informative mutation?	Ref. allele / alt. allele counts
13	Colon	0.96	0.98	100% (77378/77378)	*KRAS* p.G12V	Yes	92/56
18	Melanoma	0.85	0.92	100% (68610/68610)	*BRAF* p.V600K	Yes	119/229
19	Melanoma	0.97	0.99	100% (77833/77833)	*BRAF* p.V600K	Yes	412/256
34	Colon	0.96	0.99	100% (77055/77055)	None	NA	NA
37	Lung	0.97	0.99	100% (77645/77645)	*KRAS* p.G12V	Yes	156/418
38	Lung	0.97	0.99	100% (78015/78015)	None	NA	NA
41	Lung	0.91	0.95	100% (73427/73427)	None	NA	NA
43	Melanoma	0.94	0.97	100% (75823/75823)	*NRAS* p.Q61R	Yes	27/6

Genome Res. 2013 Mar 11.

Useful applications of smMIP

Detection and quantitation of of low-frequency mutations/ variation that is relevant:

- Human cancers: genetic heterogeneity.
- Pre-existence of subclonal resistance mutations to therapy.
- Somatic mosaicism detection.
- Noninvasively assay fetal DNA from maternal plasma at clinically relevant sites.
- *etc. ...*

Thank You!