#2640

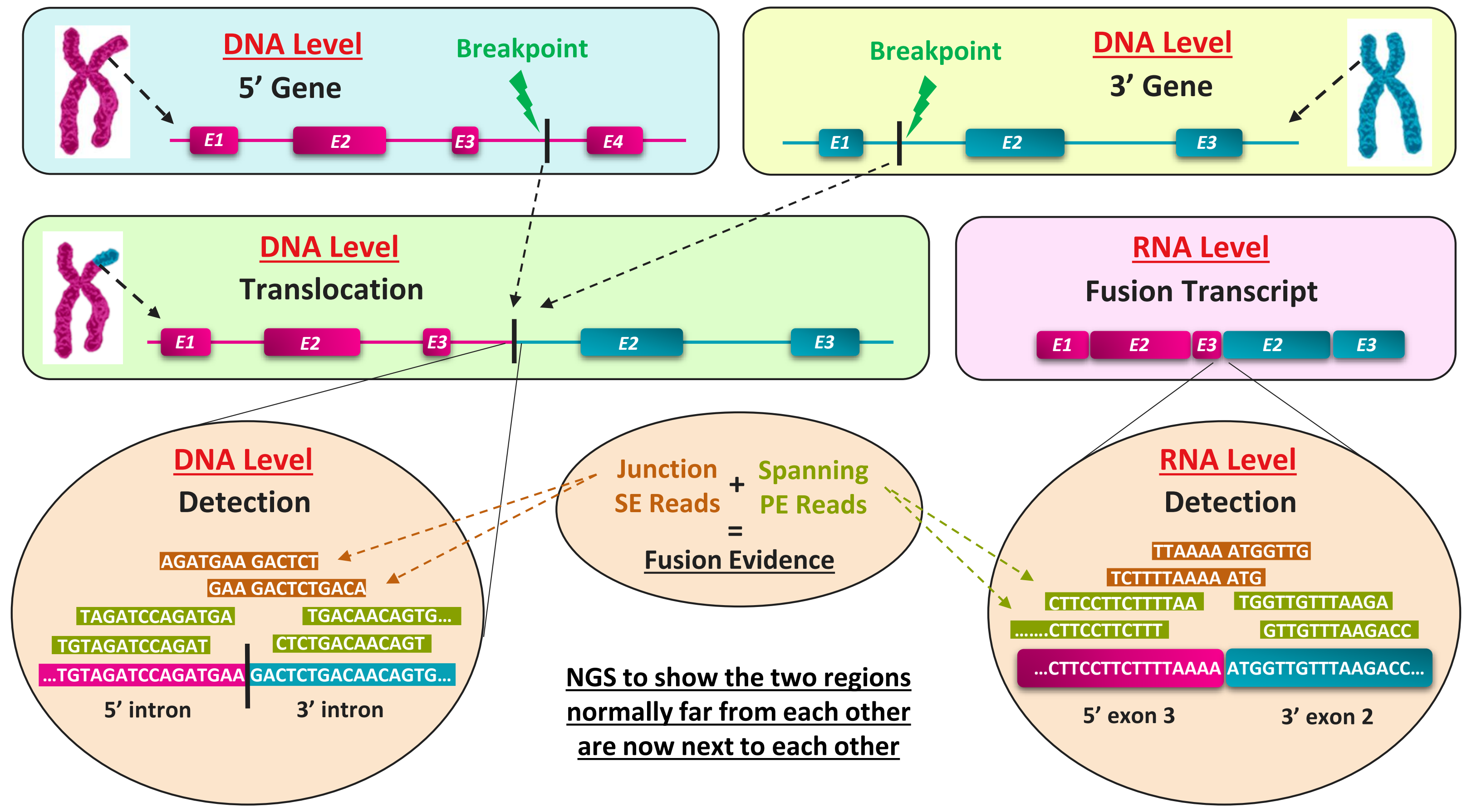
An Efficient Probe Design Algorithm For Direct Fusion Targeting From RNA

Christophe N. Magnan • Steven Lau-Rivera • Fernando Lopez Diaz • Chenyin Ou • Brad Thomas • Hyunjun Nam • Lawrence M. Weiss • Segun C. Jung • Vincent A. Funari
NeoGenomics Laboratories, Aliso Viejo, CA, Carlsbad, CA, and Houston, TX

Introduction

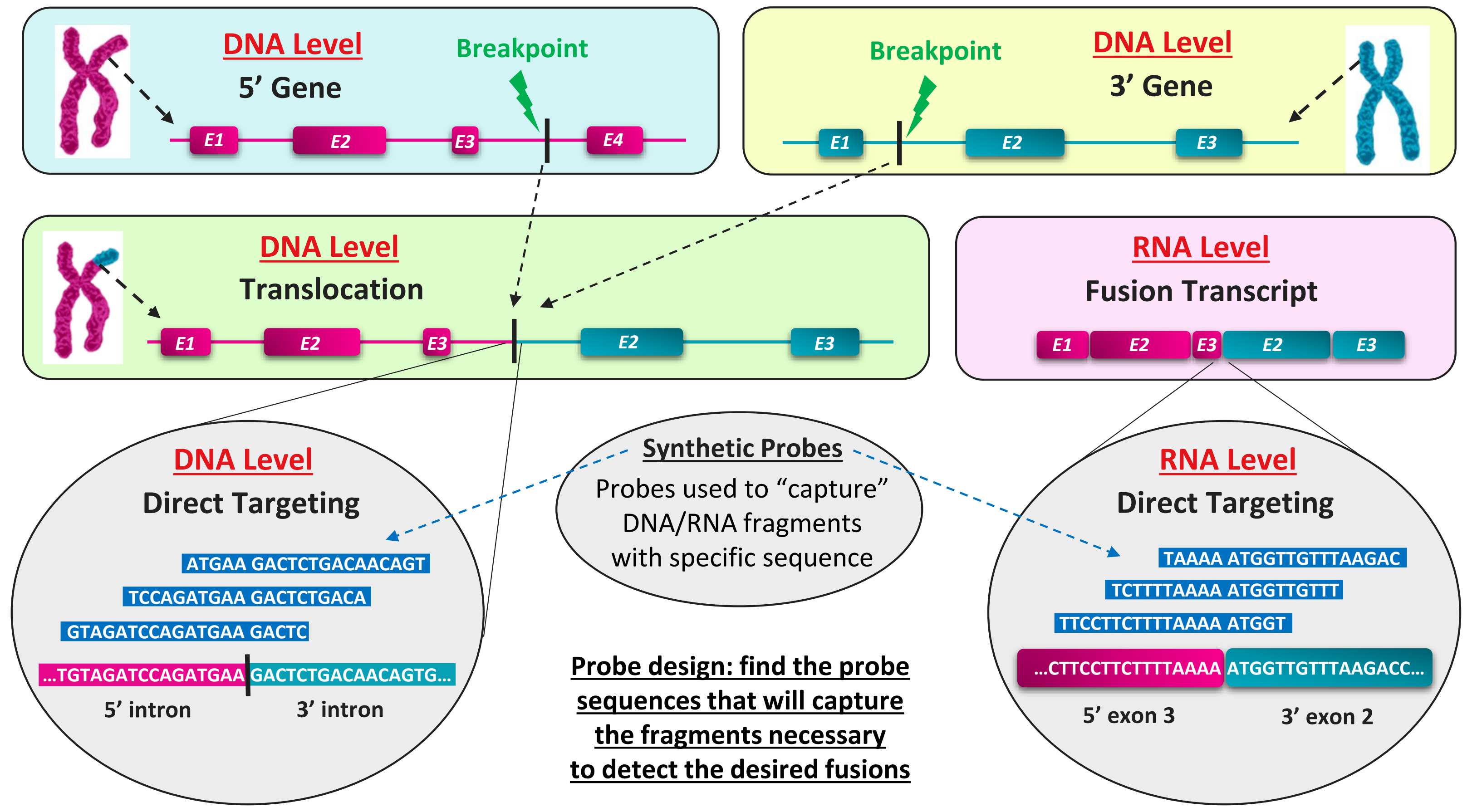
The use of sequencing technologies to detect gene fusions (GFs) from RNA shows promising results for the future of cancer diagnosis and treatment. Major obstacles for this approach include target design and lack of well-curated databases of RNA breakpoints. Currently, off-the-shelf designs include full transcript targeting that results in massive and costly amounts of data, most of which being wildtype sequences not helping the detection of GFs. Directly targeting the known GFs from RNA by designing probes directly targeting the fusion junction sequence is studied here as an alternative to whole-exome sequencing (WES). We present notably a novel algorithm capable of designing the probes to accurately target the desired fusions from RNA. The algorithm takes in the input the genomic breakpoint positions from a known gene fusion detected either from RNA or from DNA (without the source being provided to the algorithm) and outputs the genomic and transcriptomic breakpoint positions where the fusion will most likely be observed from RNA as well as the corresponding probe sequences to be synthesized for targeting of the known fusion. We show here that despite being a non-exhaustive approach, the synthesized probes successfully enrich the datasets in fusion supporting reads allowing not only a more sensitive detection of the targeted GFs but also significantly higher confidence levels in the fusion calls thanks to the increase in the number of chimeric reads used as evidence of the fusion event. Note also that this approach allows the ability to detect novel, non-targeted, fusions whenever a breakpoint is shared with one of the targeted GF.

Fusion Detection From NGS



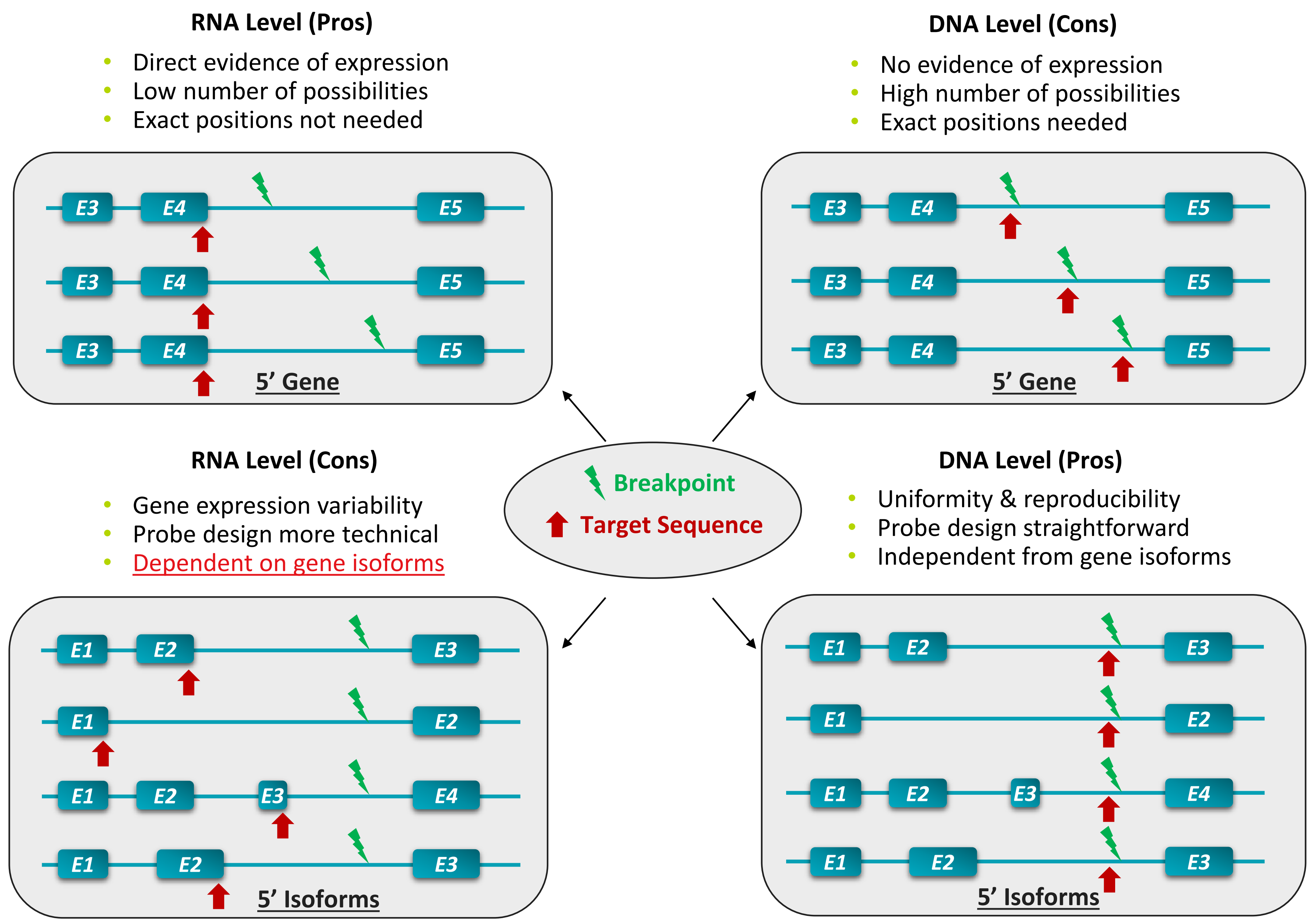
NGS to show the two regions normally far from each other are now next to each other

Probe Design Overview



Probe design: find the probe sequences that will capture the fragments necessary to detect the desired fusions

Fusion Detection from RNA vs DNA



RNA Level (Pros)

- Direct evidence of expression
- Low number of possibilities
- Exact positions not needed

DNA Level (Cons)

- No evidence of expression
- High number of possibilities
- Exact positions needed

RNA Level (Cons)

- Gene expression variability
- Probe design more technical
- Dependent on gene isoforms

DNA Level (Pros)

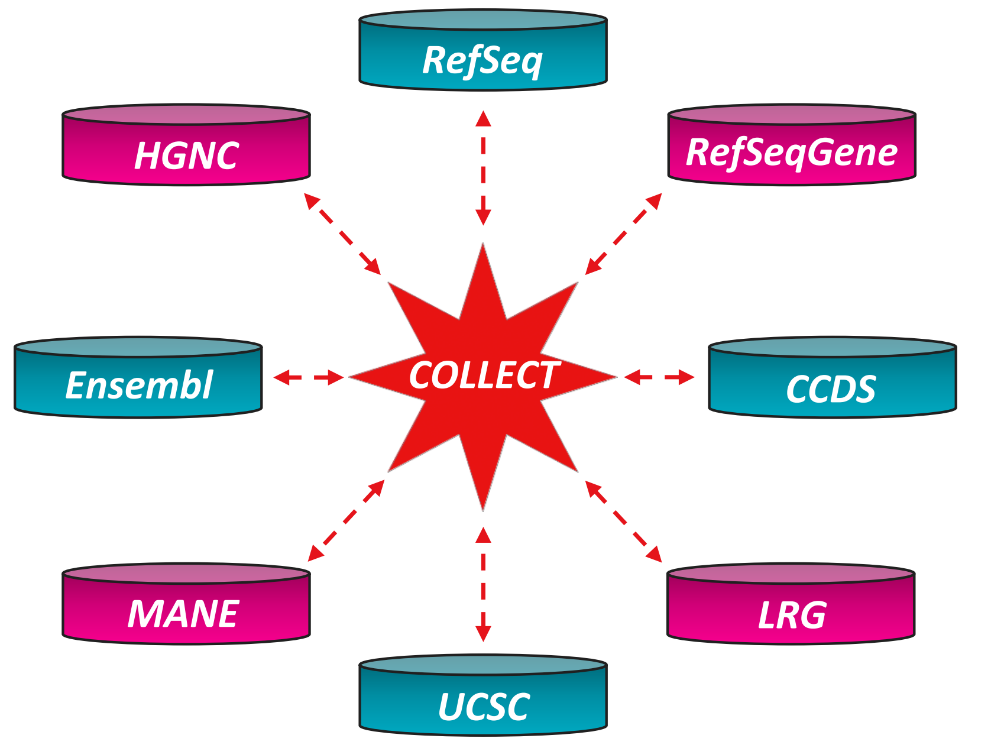
- Uniformity & reproducibility
- Probe design straightforward
- Independent from gene isoforms

Probe Design Algorithm

The challenge to design probes for direct targeting of known fusions from RNA comes from the multiple genomic and transcriptomic locations where a known fusion can be observed from RNA. A simple review of gene annotation databases like Ensembl reveals that genes can have up to 200 different known isoforms. Designing probes for every pair of isoforms for each known fusion between two genes would result in an unnecessarily large set of probes capturing massive amounts of wild-type molecules and a small number of supporting reads to detect the fusion similarly to a WES approach. Since the prevalence of each isoform in the population is not readily available in gene annotation databases, we designed an algorithm selecting one or two candidate gene isoforms for each gene involved in a known fusion based on the isoform popularity in gene annotation databases, the quality of the gene annotations, the evidence level for each isoform, the various reference sequence databases such as LRG or MANE, and the original breakpoint positions observed for the fusion previously detected from RNA or from DNA. The algorithm indirectly attempts to assess the most prevalent isoform to maximize the probability to capture the fusion with the corresponding probes.

Step 1 - Data Collection

Input Known Fusion:	CD74	5	-	150,404,680	NRG1	8	+	32,727,949



- **Gene-level information**
  - HUGO symbol, synonyms
  - Gene ID used by each database
  - Category (coding/non-coding, pseudo)
- **Isoform-level information**
  - List of known gene isoforms
  - Transcript ID used by each database
  - Isoform-specific coding information
  - Transcript, coding, protein sequences
  - Exon sequences and positions on hg38
  - Review status, evidence level
  - Available crosslinks between databases

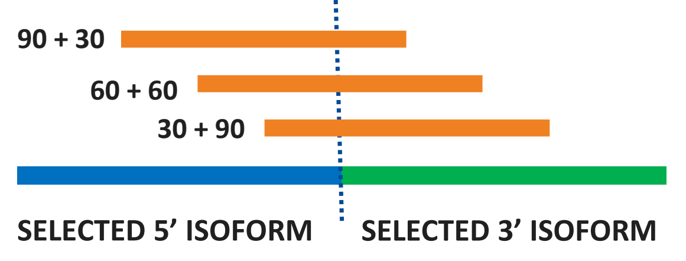
Step 2 - Data Extraction

Genomic		Transcript				Coding		Protein	
g.150404687	A	t.562	A	E5	CDS	c.555	A	p.185	Pro
g.150404686	C	t.563	C	E5	CDS	c.556	C	p.186	Pro
g.150404685	C	t.564	C	E5	CDS	c.557	C	p.186	Pro
g.150404684	G	t.565	G	E5	CDS	c.558	G	p.186	Pro
g.150404683	A	t.566	A	E5	CDS	c.559	A	p.187	Lys
g.150404682	A	t.567	A	E5	CDS	c.560	A	p.187	Lys
g.150404681	A	t.568	A	E5	CDS	c.561	A	p.187	Lys
g.150404680	G	t.569	G	E5	CDS	c.562	G	p.188	Glu
g.150402625	A	t.570	A	E6	CDS	c.563	A	p.188	Glu
g.150402624	G	t.571	G	E6	CDS	c.564	G	p.188	Glu
g.150402623	T	t.572	T	E6	CDS	c.565	T	p.189	Ser
g.150402622	C	t.573	C	E6	CDS	c.566	C	p.189	Ser
g.150402621	A	t.574	A	E6	CDS	c.567	A	p.189	Ser
g.150402620	C	t.575	C	E6	CDS	c.568	C	p.190	Leu
g.150402619	T	t.576	T	E6	CDS	c.569	T	p.190	Leu

- **Sequence-level information**
  - gdot ↔ tdot ↔ cdot ↔ pdot positions
  - Percentage identity with hg38
  - Percentage visible on hg38
  - Sequence translation and completion
- **Isoform-specific information**
  - Location of original breakpoint
  - Position breakpoint will be observed
- **Database-level information**
  - Missing crosslinks

Step 3 - Scoring & Isoform Selection

- **Isoform-level score:**
  - Isoform is in RefSeq **+1**, is reviewed **+2**
  - Isoform is RefSeq Select **+3**
  - Isoform is MANE Select **+2**, MANE Plus **+1**
  - Isoform found in LRG **+3**, in CCDS **+2**
  - TSL = 1 **+3**, TSL = 2 **+2**, TSL = 3 **+1**
  - Sequence is complete **+2**
  - 100% sequence identity with hg38 **+3**
  - >98% sequence identity with hg38 **+2**
  - >90% sequence identity with hg38 **+1**
  - No protein break **+2**, no unknown base **+2**
  - Isoform & input breakpoint positions identical **+2**
- **Isoform pair selection:**
  - Pair score = score 5' isoform + score 3' isoform
  - Ranked by decreasing score
  - Redundant lower-scoring pairs removed
  - Top scoring pair of isoforms selected
  - In specific cases, 2 pairs of isoforms selected
- **Probes designed using selected isoforms:**



Experimental Results

Two sets of probes extracted with this protocol respectively targeting 524 known gene fusions (columns 524 TF (A) and 524 TF (B) in Table 1) and 1632 known gene fusions (columns 1632 TF (A) and 1632 TF (B) in Table 1) were synthetized and tested both on a control library (SeraSeq 0710-0496) and 10 clinical samples with a known gene fusion detected using an orthogonal technology. The Agilent SureSelect Human All Exon V6 capture kit was used to compare targeting efficiency against a WES approach. The resulting number of supporting reads per fusion per million reads is reported in Table 1. Targeted enrichment of the SeraSeq control showed a 5 to 20 fold increase in supporting evidence over WES. On the 10 clinical samples, we observed a 10 to 30 fold increase in supporting reads depending on the number of targeted fusions. A higher sensitivity is observed in both cases.

Sample	Known Fusion	WES (A)	WES (B)	524 TF (A)	524 TF (B)	1632 TF (A)	1632 TF (B)
SeraSeq	CCDC6→RET	8	10	270	355	81	78
0710-0496	CD74→ROS1	35	81	592	832	246	250
	EGFR→SEPTIN14	18	17	234	315	91	80
	FGFR3→BAIAP2L1	14	5	428	326	125	72
	FGFR3→TACC3	23	9	861	879	270	203
	LMNA→NTRK1	23	11	215	280	71	67
	PAX8→PPARG	29	19	193	246	66	62
	SLC34A2→ROS1	10	22	176	425	89	142
	SLC45A3→BRAF	11	15	433	420	141	105
	TFG→NTRK1	35	40	275	377	128	132
	TMPRSS2→ERG	0	0	559	348	170	79
	TPM3→NTRK1	15	23	246	359	106	117
Avg. SeraSeq	**12 Fusions**	**18**	**21**	**373**	**430**	**132**	**116**
Clinical S1	EML4→ALK	25	15	344	NA	81	NA
Clinical S2	EWSR1→FLI1	57	38	514	NA	111	NA
Clinical S3	TES→MET	20	30	115	NA	75	NA
Clinical S4	EZR→ROS1	4	31	1,059	NA	266	NA
Clinical S5	SDC4→ROS1	5	5	3,354	NA	1,082	NA
Clinical S6	SH3BP5→PPARG	0	0	4	NA	1	NA
Clinical S7	H2BC21→NTRK1	1	1	18	NA	4	NA
Clinical S8	COL1A1→PDGFB	209	250	5,530	NA	2,289	NA
Clinical S9	KIF5B→RET	24	24	437	NA	174	NA
Clinical S10	POC1B→GLI1	15	8	416	NA	161	NA
Avg. Clinical	**10 Fusions**	**36**	**40**	**1,179**	NA	**424**	NA

Table 1: Average number of supporting reads per fusion per million reads for WES & direct targeting protocols.

Conclusion

We developed a novel algorithm capable of accurately identifying the most likely location a known fusion will be observed on RNA and automatically generating the probe sequences for oligo synthesis. Compared to a WES approach, this method increases fusion detection sensitivity, enriches for more supporting data resulting in higher confidence fusion calls, and reduces the associated costs.